# Extensive genetic divergence between recently evolved sympatric Arctic charr morphs

Jóhannes Guðbrandsson[1,2]*, Kalina H. Kapralova[1], Sigríður R. Franzdóttir[1,3], Þóra Margrét Bergsveinsdóttir[1], Völundur Hafstað[1], Zophonías O. Jónsson[1,3], Sigurður S. Snorrason[1] and Arnar Pálsson[1,3]

[1] Institute of Life and Environmental Sciences, University of Iceland, 101 Reykjavik, Iceland
[2] Marine and Freshwater Research Institute, 101 Reykjavik, Iceland
[3] Biomedical Center, University of Iceland, 101 Reykjavik, Iceland

**Abstract**

The availability of diverse ecological niches can promote adaptation of trophic specializations and related traits, as has been repeatedly observed in evolutionary radiations of freshwater fish. The role of genetics, environment and history in ecologically driven divergence and adaptation, can be studied on adaptive radiations or populations showing ecological polymorphism. Salmonids, especially the *Salvelinus* genus that includes Arctic charr (*Salvelinus alpinus*), are renowned for both phenotypic diversity and polymorphism. Arctic charr invaded Icelandic streams during the glacial retreat (about 9,000 to 12,000 years ago) and exhibits many instances of sympatric polymorphism. Particularly well studied are the four morphs in Lake Þingvallavatn in Iceland. The small benthic (SB), large benthic (LB), planktivorous (PL) and piscivorous (PI) charr differ in many regards, including size, form and life history traits. To investigate relatedness and genomic differentiation between morphs we extracted variants from developmental transcriptome data from three of those morphs, and verified 22 variants in population samples. The data reveal genetic differences between the morphs, with the two benthic morphs being more similar and the PL-charr more genetically different. The markers with high differentiation map to all linkage groups, suggesting ancient and pervasive genetic separation of these three morphs. No marker associated fully with morph, suggesting polygenic basis of traits separating them. Furthermore, gene ontology analyses suggest differences in collagen metabolism, odontogenesis and sensory systems between PL-charr and the benthic morphs. Genotyping in population samples from all four morphs confirms the genetic separation and indicates that the PI-charr are less genetically distinct than the other three morphs. The genetic separation of the other three morphs indicates certain degree of reproductive isolation. The extent of gene flow between the morphs and the nature of reproductive barriers between them remain to be elucidated.

## 1 Introduction

Organismal diversity reflects the process of evolution and highlights the importance of natural selection in building and maintaining adaptations (Darwin 1859). While purifying selection preserves adaptations and biological functions, positive selection alters phenotypic traits and frequencies of genetic variations influencing them (Via 2001, 2009). However, the complex relationships between genes and traits has made the quest for finding the genes under selection difficult. Feeding is a primary function in all animals and in nature we see many examples of spectacular adaptive radiation where natural selection has generated a range of adaptations in specific feeding structures and foraging behavior (Losos & Ricklefs 2009; Seehausen & Wagner 2014). Recently, studies have demonstrating how positive selection can shift phenotypic distributions on generational time scales (Grant & Grant 2002; Nosil 2012). Freshwater fish exhibit great diversity in feeding specializations, e.g. in the structure and function of the feeding apparatus and head morphology (Seehausen & Wagner 2014). The recurrent ecological specializations found in many species and populations of fish, represent interesting study systems and set of functional phenotypes for analyses of evolutionary change, convergence and parallel adaptation (Muschick *et al.* 2012).

*Corresponding author – Tel: +354-575-2615 – johannes.gudbrandsson@hafogvatn.is

## 1.1 Genome wide divergence or islands of differentiation?

In the last decade the genomics revolution spawned powerful tools for studying both the connections between genetic and phenotypic variation, e.g. how variations in the form of feeding structures are caused by differential expression of key developmental genes (Abzhanov *et al.* 2006; Guðbrandsson *et al.* 2018), and to determine to what extent these differences are due to genetic differences (and how they are distributed in genomes) (Wolf & Ellegren 2016). These methods have enabled studies of the role of feedback between the organism and its environment, the role of plasticity in generating functional variation and possibly promoting adaptation (Morris & Rogers 2014; Abouheif *et al.* 2014). More central to this study, genomics can reveal differences between populations and ecotypes, and identify genes and pathways that may be under positive selection in natural populations (Malinsky *et al.* 2015; Wolf & Ellegren 2016).

The differentiation and divergence between related groups, populations, sub-species or species is influenced by genomic parameters, population genetic history and evolutionary forces (Seehausen *et al.* 2014; Vijay *et al.* 2017). Depending on multiple factors, e.g. the strength of selection, nature of the adaptive traits, genomic structure, population genetic parameters and history, the genetic separation of populations or subspecies may be rather uniform over the entire genome or localized to "genomic islands of differentiation" (Malinsky *et al.* 2015; Wolf & Ellegren 2016). Studies of closely related species and ecologically distinct populations within species can detect the genetic correlates of adaptive traits (Pease *et al.* 2016), but this is not straightforward. Positive selection can affect alleles, both from older standing variation and more recent variants, private to certain populations (Vijay *et al.* 2017). Adaptive introgression or environmental sorting of an ancestral, possibly balanced, polymorphism, and homoploid speciation can also complicate the analyses (Pardo-Diaz *et al.* 2012; Cruickshank & Hahn 2014; Guerrero & Hahn 2017). Furthermore, distinct populations of a species adapting to different environments, can be viewed as being scattered along a "speciation continuum" (Theis *et al.* 2014; Seehausen *et al.* 2014). Fluctuations in environmental or population genetic parameters can shift populations along this continuum (Grant & Grant 2008). Specifically, changes in the environment or gene flow between populations, can lead to the merging of previously distinct populations, speed up their divergence or produce reproductively isolated species (Hendry *et al.* 2009; Seehausen *et al.* 2014; Lowry & Gould 2016). We are keen to explore the genome-wide differentiation of sympatric polymorphism within a species that recently colonized new habitats, with the aim to study the genetics of ecological specializations.

## 1.2 Resource polymorphism along a benthic - limnetic axis

Variation in resource utilization among allopatric populations can lead to divergence of traits which, given time, can result in adaptive divergence among these populations. Discrete variation in resource use among individuals in the same area, e.g. foraging in different habitats, can also generate divergent traits within populations (Skúlason & Smith 1995; Bernatchez *et al.* 2010). The occurrence of discrete phenotypes (morphs) of a species living in the same area and diverging in traits that relate to utilization of different resources (food, breeding grounds, etc) has been called resource polymorphism. Typically, diverging traits of resource morphs involve morphology, behavior and/or life history characteristics (Skúlason & Smith 1995; Smith & Skúlason 1996). Resource polymorphism can arise through developmental plasticity within a homogenic population, by natural selection on genetically encoded traits or a combination of these mechanisms. For instance, the broad and narrow headed morphs of the European eel (*Anguilla anguilla*) seem to be mainly determined by environmental factors (De Meyer *et al.* 2016) and ecomorphs of killer whales (*Orcinus orca*) were hypothesized to have originated through plastic responses of a small founder population (Foote *et al.* 2016). Examples of genetically determined resource polymorphism can be found in crater lake cichlid fishes, e.g. were a single locus effects jaw and body shape (Fruciano *et al.* 2016). Likewise, in benthic and limnetic "species" of threespine stickleback (*Gasterosteus aculeatus*), the same loci show signs of differentiation (Jones *et al.* 2012).

Salmonids are renowned for their phenotypic diversity, both among and within populations, with multiple examples of resource polymorphism. The most phenotypically diverse and polymorphic species seem to be in the *Salvelinus* genus, including Arctic charr *Salvelinus alpinus* (Klemetsen 2013), Lake charr (also called Lake trout) *Salvelinus namaycush* (Muir *et al.* 2016) and Dolly Varden charr (*Salvelinus malma*) (also called Dolly Varden trout) with as many as seven morphs found in the same lake (Markevich *et al.* 2018). Arctic charr colonized lakes and rivers on the northern hemisphere after the last glaciation period (approx 9,000-12,000 years ago) (Snorrason & Skúlason 2004; Noakes 2008; Klemetsen 2010). In Iceland, multiple lakes harbour polymorphic Arctic charr (Skúlason *et al.* 1992; Snorrason & Skúlason 2004; Wilson *et al.* 2004; Woods *et al.* 2012b) and a unique small benthic morphotype is found in many streams and ponds across the country, especially in cold springs with lava-

rock bottom in the geologically younger parts of the island (Kapralova *et al.* 2011; Kristjánsson *et al.* 2012). One of four sympatric charr morphs found in Lake Þingvallavatn, Iceland's largest lake, is of this type. Population genetics show that different populations of Arctic charr in Iceland are grouped by geography, not morphotype (Gíslason 1998; Kapralova *et al.* 2011), supporting the notion that small benthic morphs and other derived forms have evolved independently in different locations. For instance, the four Lake Þingvallavatn morphs are more closely related to one another than to other charr populations (Kapralova *et al.* 2011). Genetic analysis using allozymes, microsatelites and RADsequencing in Norway, Scotland and Transbaikalia also indicate closely related sympatric morphs (Hindar *et al.* 1986; Gordeeva *et al.* 2015; Jacobs *et al.* 2018), while in other cases sympatric charr morphs seem to have emerged by more than one invasion (Verspoor *et al.* 2010).

## 1.3 The four sympatric charr morphs in Lake Þingvallavatn

Lake Þingvallavatn formed as the Icelandic ice-cap receded ∼10,000 years ago and was shaped by volcanic activity and isostatic rebound during its formation. The lake is located in a rift zone and has since been influenced by tectonic movements, causing extensive subsidence and horizontal extension with extensive rift-forming in the central graben (Saemundsson 1992). The lake harbours four distinct Arctic charr morphs: Small benthic (SB), Large benthic (LB), Planktivorous (PL) and Piscivorous (PI) charr, that differ ecologically along a benthic - limnetic axis and this is reflected in their form, size, habitat use, diet and life history characteristics (Sandlund *et al.* 1987; Jonsson *et al.* 1988; Malmquist *et al.* 1992)(Figure 1). Furthermore, the morphs differ extensively *e.g.* in their spawning times (Skúlason *et al.* 1989b) and parasite loads (Frandsen *et al.* 1989; Kapralova *et al.* 2013). The morphs represent genuine resource polymorphism as defined by (Skúlason & Smith 1995). Common garden experiments on offspring indicated heritable differences in various traits between morphs, e.g. morphology, sexual maturation rates, foraging behaviour (Skúlason *et al.* 1989a, 1993, 1996) but plasticity was also found to be significant (Parsons *et al.* 2010, 2011), which suggests that both genetic differences and plasticity influence variation between Arctic charr morphs.

The earliest population genetic studies of the Lake Þingvallavatn charr found variation in few genetic markers, and weak separation between morphs (Magnusson & Ferguson 1987; Danzmann *et al.* 1991; Volpe & Ferguson 1996). The estimated relationship between the morphs varied by studies (markers), two datasets placed SB-charr as the most distantly related morph (Magnusson & Ferguson 1987; Gíslason 1998), another placed LB-charr as the most distant morph (Kapralova 2008) while an mtDNA study clustered the benthic morphs (LB- and SB-charr) and placed PL-charr as an outgroup (Volpe & Ferguson 1996). Curiously the PI-charr clustered with both PL- and LB-charr. A study of 9 microsatellite markers in SB- and PL-charr from five spawning sites and LB-charr confirmed morph difference (average $F_{ST}$ = 0.039), but did not resolve their relationship (Kapralova *et al.* 2011). Crucially, those data and coalescence modeling were more consistent with a scenario assuming a brief initial phase in allopatry of PL- and SB-charr with subsequent coexistence, than a sympatric origin of these morphs (Kapralova *et al.* 2011). This observation was supported by data showing high genetic separation in two immunological genes (Kapralova *et al.* 2013) and substantial variation in other loci between morphs (Gudbrandsson *et al.* 2016).

We have studied the developmental and molecular correlates of morph divergence, by analyzing morphology of embryos, miRNA and mRNA expression (Kapralova *et al.* 2014; Ahi *et al.* 2014; Kapralova *et al.* 2015; Gudbrandsson *et al.* 2016; Guðbrandsson *et al.* 2018). Most recently, RNA-sequencing of embryos of three of the four morphs (excluding PI-charr), reared in a common garden set-up, revealed about 2000 transcripts with consistent expression differences between morphs (Guðbrandsson *et al.* 2018). The PL-, LB- and SB-charr differed at the transcriptome level and the results suggested a closer relationship between the two benthic morphs (LB- and SB-charr), with PL-charr more distinct. No enrichment of specific biological pathways (GO analyses) was found for the morph specific transcripts, implying that multiple genes and diverse pathways contribute to the developmental morph differences. However, qPCR analysis of candidate co-expressed genes, including genes related to extra-cellular matrix organization and skeletogenesis as well as their putative regulators, revealed variation along the benthic - limnetic axis (Ahi *et al.* 2013, 2014, 2015). Similarly, Macqueen *et al.* (2011) found differential expression of three genes in the *mTOR*-pathway in muscles of generalist vs. small benthic morphs from multiple locations within Iceland (Macqueen *et al.* 2011). Notably, we found the same increased expression of one of these genes, *Eukaryotic translation initiation factor 4E binding protein 1 (Eif4ebp1)*, in SB-charr from Lake Þingvallavatn (Guðbrandsson *et al.* 2018).

Like other freshwater fishes (Seehausen & Wagner 2014; Bernatchez *et al.* 2010; Berthelot *et al.* 2014), Arctic charr has utility for studies of recent adaptations and the genetics of evolutionary change (Snorrason & Skúlason 2004). Relatively few genomic resources were available for Arctic charr until recently, when a reference genome, that will aid further genomic research of the species, was published (Christensen *et al.* 2018). Two other salmonid

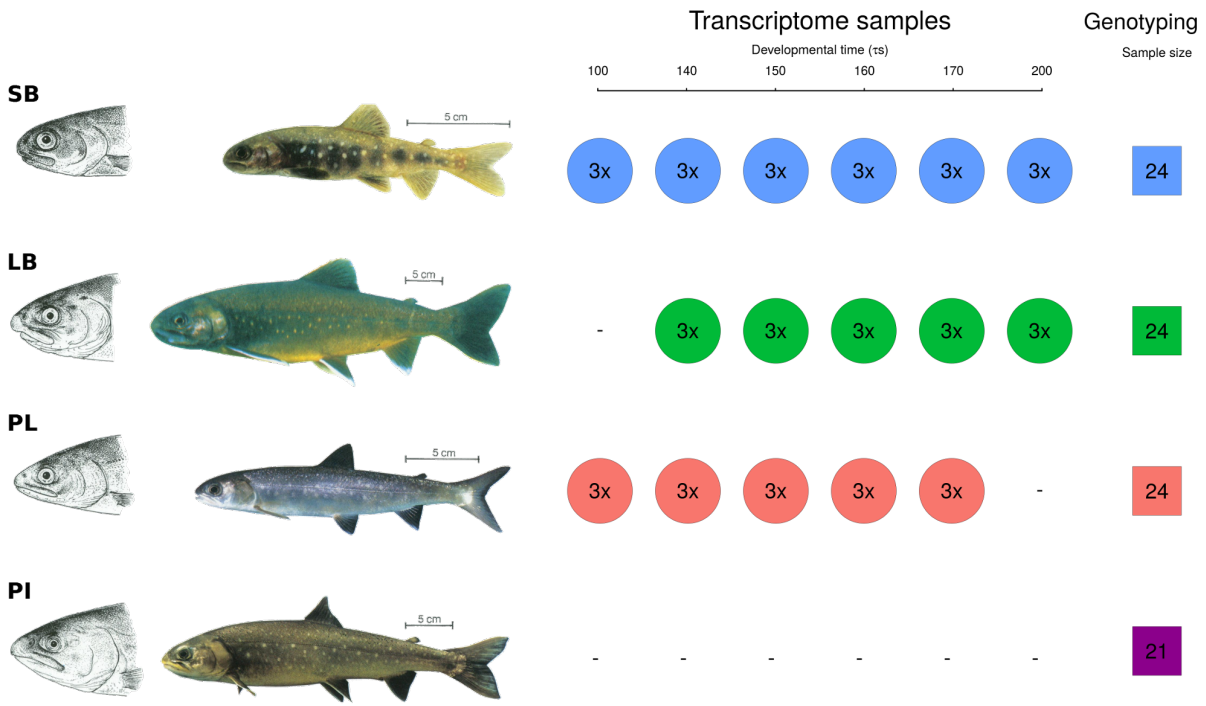


Figure 1: The phenotypically distinct sympatric Arctic charr from Lake Þingvallavatn and the sampling strategy. **A)** The four sympatric morphs are Small benthic (SB), Large benthic (LB), Planktivorous (PL) and Piscivorous (PI) charr. They differ in size (size bars = 5 cm), head and feeding morphology and pigmentation. Adapted from Sandlund *et al.* (1992) ©Wiley-Blackwell, drawings by Eggert Pétursson. **B)** Sampling of charr for transcriptome sequencing (circles) and genotyping (squares) of population samples. The top 3 morphs were mined for genetic variation in the transcriptome and population samples were studied from all four, to confirm genetic variants. The transcriptome samples came from embryos at 6 developmental stages prior to hatching, from $100\tau s$ to $200\tau s$, in the three morphs (circles)(Guðbrandsson *et al.* 2018). Sampling of each morph and developmental timepoint combination was replicated three times (biological replicates), each sample being a pool of mRNA from three embryos. Six timepoints were sampled of SB-charr, and five of LB- and PL-charr embryos. The population samples (squares) were obtained by gill netting on the spawning grounds, see methods. The morph coloring scheme (SB: blue, LB: green, PL: red and PI: purple) will be retained throughout the manuscript.

genomes are published (Berthelot *et al.* 2014; Lien *et al.* 2016), and a third (Coho salmon, *Oncorhynchus kisutch*) is available on NCBI (Genebank assembly GCA_002021735.1). About 88–103 million years ago an ancestor of Salmonids underwent a whole genome duplication (Ss4R), the fourth on the vertebrate lineage (Allendorf & Thorgaard 1984; Macqueen & Johnston 2014; Berthelot *et al.* 2014). Comparisons of several salmonids established significant synteny of their linkage groups (Danzmann *et al.* 2005; Sutherland *et al.* 2016), despite rearrangements (Timusk *et al.* 2011; Nugent *et al.* 2017). Thus a substantial fraction of salmonid genes are paralogs, and are found in syntenic regions (Nugent *et al.* 2017; Christensen *et al.* 2018). This understandably complicates analyses of differential gene expression (Gudbrandsson *et al.* 2016; Guðbrandsson *et al.* 2018) and genetic variation.

Here we mined the developmental transcriptome (Guðbrandsson *et al.* 2018) of three of the morphs (LB-, SB- and PL-charr) to test for genetic differences between them, elucidate their evolutionary relationship and identify the loci and developmental pathways of morph differentiation. The data were used to evaluate three hypotheses about the causes of morph separation:

I Because the salmonid's homing behavior leads offspring to spawn in the same location as their parents, heterogeneity in spawning places and micro-environments can lead to environmentally induced gene expression and phenotypic differences. In this scenario morphs are environmentally induced, and genetic differences between them minor.

II Ecological selection on specific traits has lead to genetic separation of morphs, seen as allele frequency dif-

ferences at variants in key genes related to fitness traits, in the face of gene flow between morphs (Seehausen *et al.* 2014; Wolf & Ellegren 2016). Under this model, genomic island of differentiation are expected, with limited genetic differences between morphs in the rest of the genome.

III Pre-zygotic barriers such as spatial and temporal separation in spawning between morphs (Skúlason *et al.* 1989b) or behavioral/mating differences have kept the morphs reproductively isolated for some time (as has been indicated for SB- and PL-charr (Kapralova *et al.* 2011)). Under this scenario, modest genetic separation among the morphs is expected, on all linkage groups.

These three hypotheses are not mutually exclusive, and some may apply to specific pairs of morphs. The data enabled analyses of the genetic separation of three of the sympatric morphs, the origin of the rare piscivorous charr, the genomic patterns of differentiation between morphs and the genes and molecular systems that may associate with their specializations.

# 2 Materials and Methods

## 2.1 Sampling for transcriptome and population genetics

The fishing and sampling of embryos from Lake Þingvallavatn for developmental transcriptome analysis is described in Guðbrandsson *et al.* (2018). Briefly, we caught running SB-, LB- and PL-charr, made crosses and reared embryos in a common garden set-up at the Hólar Aquaculture station, Verið (Sauðárkrókur). Samples of embryos were taken at 5-6 developmental timepoints depending on morph (see Figure 1). For each morph and developmental timepoint three biological replicates were analyzed, each of those isolated from pools of 3 embryos. The total number of samples was 48 (Figure 1). The crosses were made using multiple parents. For most samples the embryos came from crosses created in the 2010 spawning season (SB 150- 200$\tau s$, PL 140-170 $\tau s$, LB 140-200$\tau s$). For SB- and PL-charr ten males and ten females were used for the cross but five of each sex for LB. Due to poor RNA quality of some samples from early developmental stages, we added samples from the 2011 spawning season. For timepoint 100$\tau s$ in PL we used a cross from the 2011 spawning season with the same setup as before. For SB at timepoints 100$\tau s$ and 140$\tau s$ we used two single parent crosses due to difficulties of finding running SB-charr. Samples SB100A and SB100B came from one cross and SB100C and all three samples for timepoint 140$\tau s$ came from the other one.

For verification of single nucleotide polymorphism's (SNP's) and comparisons between morphs we sampled 93 sexually mature (most of them running) individuals of the four morphs; 24 SB-, 24 LB-, 24 PL- and 21 PI-charr. The fish were caught by gillnet fishing in Ólafsdrattur and Mjóanes in 2015 and 2016.

As before (Guðbrandsson *et al.* 2018), all fishing in Lake Þingvallavatn was with permissions obtained from the farmers in Mjóanes and from the Þingvellir National Park commission. Ethics committee approval is not needed for regular or scientific fishing in Iceland (The Icelandic law on Animal protection, Law 15/1994, last updated with Law 55/2013).

## 2.2 RNA isolation and sequencing

Total mRNA was isolated from a pool of three embryos for each sample. The RNA was quality checked and used as a template for 48 cDNA libraries subjected to Illumina sequencing. The sequencing was performed on Hiseq2000 at deCODE genetics (Reykjavík, Iceland) yielding 1,188 million 101 paired end reads. The sequencing reads from the 48 samples were deposited into the NCBI SRA archive under BioProject identifier PRJNA391695 and with accession numbers: SRS2316381 to SRS2316428. For more detailed description about RNA isolation, quality checks and library construction see Guðbrandsson *et al.* (2018).

## 2.3 Transcriptome assembly, Abundance estimation and Annotation

The transcriptome assembly and annotation were described previously (Guðbrandsson *et al.* 2018), but briefly the reads were quality trimmed and adapters removed using Trim Galore! (version 0.3.3, Krueger 2012). The filtered reads from all samples were assembled with Trinity (version v2.1.0, Grabherr *et al.* 2011). We used kallisto (version v0.42.4, Bray *et al.* 2016) to estimate transcripts abundance. To speed up the annotation process transcripts with fewer than 200 mapped reads were not retained for annotation, another reason being that lowly expressed transcripts are unlikely to supply variants with enough coverage to surpass quality filters. The

transcripts were annotated with the Trinotate pipeline (version 2.0.2, Haas 2015). Orthologs of the transcripts in salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*) mRNA and protein sequences were found using `blastn` and `blastx` respectively (Altschul *et al.* 1990). Annotation from NCBI *Salmo salar* Annotation Release 100 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Salmo_salar/100/, retrieved 2015-12-17) and SalmoBase (Samy *et al.* 2017, http://salmobase.org/, version from 2015-09-18) were both used for salmon and annotation for rainbow trout came from Berthelot *et al.* (2014, http://www.genoscope.cns.fr/trout/data/, version from 2014-05-19). Only the best match was retained for each reference database. For further details about the annotation process and parameters we refer to Guðbrandsson *et al.* (2018).

## 2.4 Variant calling and quality filtering

To identify genetic variation in the transcriptome we mapped the quality trimmed reads to the complete Trinity assembly (annotated and non-annotated transcripts). Pseudobam files generated by kallisto were supplied to eXpress (version 1.5.1 Roberts & Pachter 2012) to get single alignment for multi-mapping reads. eXpress uses expectation–maximization (EM) algorithm to get posterior probabilities for read placements and assigns reads to transcripts by the posterior probability. We used the default eXpress parameters except we set the batch option to 10 to get more EM-rounds and better assignment of reads. Reads with more than 10 mismatches were identified and removed using samtools (version 1.1, Li *et al.* 2009) and bamtools (version 2.3.0, Barnett *et al.* 2011).

Candidate variants (hereafter variants) were called with FreeBayes (version v1.0.1-2-g0cb2697, Garrison & Marth 2012) on all the samples simultaneously. The coverage threshold was set to 480 reads and the threshold for a variant allele to be called to 48 reads. The other options used for FreeBayes were; `-use-duplicate-reads`, `-ploidy 6`, `-use-best-n-alleles 4` and `-pooled-discrete`. We only processed bi-allelic variants further. Variants were filtered based on coverage and allele frequency. We only consider positions with minimum coverage of 10 reads in at least 30 samples, thereof the minimum of eight samples in each morph. Minimum allele frequency of the alternative allele was set to 10% and the maximum to 70%.

Due to the genome duplication in the salmonid ancestor (Allendorf & Thorgaard 1984; Moghadam *et al.* 2011) some of the candidate variants might reflect sequence divergence of paralogous genes rather than true variants. We set out to remove such candidates by identifying variants with small differences in allele frequency between samples (and thus individuals). Each sample represents 3 individuals, with 6 copies of each chromosome. True biallelic segregating markers will therefore only have seven possible genotype combinations (0:6, 1:5, 2:4, 3:3, etc). This should result in variable allele frequency among samples as it is unlikely that the same combination of genotypes will be pooled in all incidents. In contrast, candidate variants due to fixed differences in paralogous genes will have similar allele frequency in all samples. To estimate this we calculated hierarchical F-statistics (Hartl & Clark 2006, p. 280-283) in R (R Core Team 2015). First we estimated the variation between samples, with a statistic termed $F_{PT}$; were $P$ stands for pool (each sample is a pool of three individuals) and $T$ the total population. We also calculated a traditional $F_{ST}$, to capture variation in allele frequencies between morphs (subpopulations). Note, because of the sampling of related individuals the values of $F_{ST}$ are not comparable to other studies (see results).

The $F_{PT}$ statistic is analogous to the classical $F_{IT}$, so low $F_{PT}$ values indicate low variation between samples just as low $F_{IT}$ values indicate low variation among individuals (excess heterozygosity). Filtering on $F_{PT}$ is therefore similar to removing markers that deviate from Hardy-Weinberg equilibrium when genotypes are on individual basis. To determine a reasonable cutoff for $F_{PT}$ we simulated 288 chromosomes with eight different alternative allele frequencies (0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7). Chromosomes were distributed randomly among samples with the same experimental setup as in our study and hierarchical F-statistics were calculated. For each alternative allele frequency 10.000 replicates were simulated. Based on this simulation (see figure S1) we chose $F_{PT}$ = 0.1 as a cutoff and all variants below this value were removed. Samples with coverage of five reads or less were not included in the calculations of F-statistics.

We ran FreeBayes again on filtered variants in order to phase variants into longer haplotypes, when possible. We used a haplotype length window of 70 basepairs (bp) and only asked for biallelic variants (`-use-best-n-allels 2`). Nevertheless some variants in the output were not biallelic and therefore removed. We also removed variants if the new estimate of $F_{PT}$ was the below the cutoff (0.1) in the new run. No further filtering of variants was performed.

## 2.5 Analysis of variant functionality and distribution among morphs

Open reading frame prediction with Transdecoder (Haas & Papanicolaou 2015) in the Trinotate pipeline (see above) was used to estimate the position (UTR, CDS) of the variants and determine if they affected the protein sequence. Further statistical analysis was performed in the R environment (R Core Team 2015) using the VariantAnnotation package (Obenchain *et al.* 2014) to handle the variant data. The alternative allele frequency (based on read counts) for each sample was used for further analysis of genetic separation between groups. To study the distribution of variation among samples we did principal component analysis (PCA) with the built in prcomp-function in R. Missing values in the dataset were populated with the mean alternative allele frequency for the variant (21087 out of 1848192 values were missing or 1.14%), prior to the PCA analyses. We calculated the mean allele frequency of each variant for each morph, and also the deviation from the other two as the sum of allele frequency difference between them. For example the deviation for PL is $d_{PL} = d(PL, SB) + d(PL, LB)$ where $d(PL, SB)$ is the difference in mean allele frequency for PL- and SB-charr. To screen for morph separation, we calculated $F_{ST}$ comparing allele frequency by morphs (see above). Variants with $F_{ST}$ above 0.2 were analyzed for gene ontology (GO) enrichment. The GO tests were performed with the goseq-package in R (Young *et al.* 2010), separately for variants with the largest deviation of mean allele frequency for each morph. The goseq-package accounts for transcript length as more variants are expected in longer transcripts. All transcripts with variants (8,961 transcipts) were used as the reference set for GO-enrichment tests. We also ran GO tests on the gene level using annotation to Salmobase. Gene length was not taken into account in that case and the reference set consisted of 7317 genes. GO categories were also mapped to their ancestor using the GO.db-package in R (Carlson 2015). We only tested biological processes, omitting categories closer than 3 steps from the root of the annotation tree. The gene ontology annotation was based on SalmoBase, see above and Guðbrandsson *et al.* (2018). Categories with false discovery rate (Benjamini & Hochberg 1995) below 0.01 were considered significant. We used 0.01 instead of the classic 0.05 as six tests were conducted (each morph on transcript and gene level). We cataloged variants private to one morph, the criterion being less than 1% frequency of the alternative allele in the other morphs. Morph specific lists of private alleles were analyzed for GO enrichment the same way.

## 2.6 Genomic distribution of candidate variants

With the Arctic charr reference genome (Christensen *et al.* 2018, Assembly `GCA_002910315.2 ASM291031v2`) it became possible to assign variants to linkage groups. A sequence 200 bp upstream and downstream of each variant in the transcriptome contigs were mapped to the genome with `blastn` within R (Hahsler & Nagar 2017) using: `-max_target_seqs 2 -max_hsps 12 -culling_limit 2`. In the case of more than one blast hit for each sub-sequence, hits with the highest bit score that included the variant were chosen. If no hit included the variant the hit with highest bit score was used. If more than one hit was equally likely, hits to chromosomes were priorities to hits to contigs or scaffolds. If equally likely hits mapped to the same chromosome or scaffold in the genome and were within 50kb from each other the hit with the lower position was chosen. The remaining variants were left unplaced.

## 2.7 Genotyping of candidate variants in a population sample

We genotyped 93 adult fish from the four morphs, PI- (21) PL- (24), LB- (24) and SB-charr (24). DNA was extracted with phenol chloroform according to standard protocols. Candidate markers were chosen based on high $F_{ST}$ values and predicted biological functions of the genes they affected. To design primers for genotyping we aligned regions around candidate variants, contigs from the assembly and similar regions in *O. mykiss* and *S. salar* (retrieved by `blast`). Locations where updated when the *S. alpinus* genome became available (Christensen *et al.* 2018). Markers were also chosen to tag independent linkage groups, but a handful was picked to survey variation in specific chromosomal regions. Sequences surrounding 23 variants (Table S5) were submitted to LGC genomics which designed KASP TM primers (He *et al.* 2014). The reactions were run on an Applied Biosystems 7500 Real-Time PCR machine, with the standard KASP program. All but one set of primers (*mrpl52_T76A*) passed test runs, and were run on 93 samples (with 3 blank controls). Analyses of genotype and allele frequencies, correspondence of genotypes to Hardy Weinberg proportions, and the correlation of allele frequencies in the transcriptome and charr populations (PL, LB and SB-charr) were conducted with R packages (`pegas, adegenet` and `hierfstat`) and custom made scripts (Jombart 2008; Paradis 2010; Goudet 2005).

# 3 Results

## 3.1 Genetic variation separating three sympatric Arctic charr morphs

To estimate the relatedness of sympatric charr, to study genome wide patterns of differentiation and to look for candidate genes related to morph separation, we screened for genetic variation in developmental transcriptomes of three Þingvallavatn morphs (SB, LB and PL-charr) (Guðbrandsson *et al.* 2018). Because of the extra whole genome duplication in the ancestor of salmonids (Allendorf & Thorgaard 1984; Moghadam *et al.* 2011) and as each sample was a pool from three individuals, we developed a F-statistic based filter ($F_{PT}$, see methods) to remove spurious variants caused by sequence divergence of paralogous genes. Variants with similar frequency in all samples most likely reflect sequence differences between stably expressed paralogs, but true polymorphisms should differ in frequency among samples. Simulations confirmed this assumption (see Figure S1), and from them we decided on $F_{PT} = 0.1$ as a threshold. It should be noted that paralogs differing in expression levels between morphs and/or timepoints escaped this $F_{PT}$-filter and some variants in the dataset could be of that nature. Thus a subset of variants was subject to validation (see below). After filtering, 19,252 variants remained, in 8,961 transcripts of 7,968 genes (Tables 1 and S1 and Supplementary file S2). As the data came from genetically related samples (each sample a pool of 3 embryos, from families produced by multiparent crosses), we could not apply standard population genetic analyses. Instead, we conducted principal component analyses, calculated $F_{ST}$ between groups, tested for GO enrichment and mapped variants to linkage groups in order to characterize the patterns of genetic variation.

Table 1: The number of variants after each filtering step in the variant-calling pipeline.

| | Variants |
|---|---|
| Freebayes first run | 143,744 |
| Cov and freq filters | 59,292 |
| $F_{PT}$ filter | 23,408 |
| Freebayes second run | 19,575 |
| Final set | 19,252 |

The three morphs separated at the genetic level. The first principal component (calculated for all variants) distinguished the planktivorous morph (PL-charr) from the two benthic morphs (LB- and SB-charr), which in turn separated on the second axis (Figure 2). The first two axes explain 19.5% and 12.7% of the variation, and other axes 4.2% at the most each (data not show). To further investigate the genetic difference between morphs we studied variants with high $F_{ST}$-values and private alleles in each morph (note that due to the sequencing of pooled samples from sibling embryos, the $F_{ST}$ could be inflated). Using $F_{ST} = 0.2$ as a cutoff yielded 2331 variants (Figure 3). Variants were categorized by the magnitude of the absolute allele frequency difference between one morph and the other two, and called LB-, SB- or PL-charr specific variants. For instance, the SB-specific variants, differed most strongly in allele frequency in SB-charr versus the average allele frequency of LB- and PL-charr. A significant excess ($\chi^2$=140.3, df = 2, $p < 0.0001$) of variants (1174) belonged to the PL-specific category (separating PL-charr and the two benthic morphs), while 605 and 552 variants associated with SB- and LB-charr, respectively.

A more stringent $F_{ST}$ cut-off (0.4) exacerbated the PL-charr vs. benthic division (Figure S4). A higher fraction of variants ($\chi^2$=56.6, df=2, $p > 0.0001$) separated PL-charr from the other two (202 PL-specific, 71 SB-specific, and 51 LB-specific variants). The same pattern was observed for private variants (alleles with frequency <1% were considered absent from a morph), 56, 18 and 13 were private to PL-, SB- and LB-charr, respectively (Table S3). Note, differences in allele frequencies were between groups, and not evolutionarily polarized (ancestral versus derived). For instance, PL-specific variants reflect either higher frequency of derived alleles in the PL-charr, or in both benthic morphs. Based on the PC-analyses, $F_{ST}$ values and private alleles, we concluded that the largest genetic separation of the sympatric charr in Lake Þingvallavatn was between PL-charr and the two benthic morphs. Next we investigated the patterns of differentiation, both in biological function and chromosomal distribution.

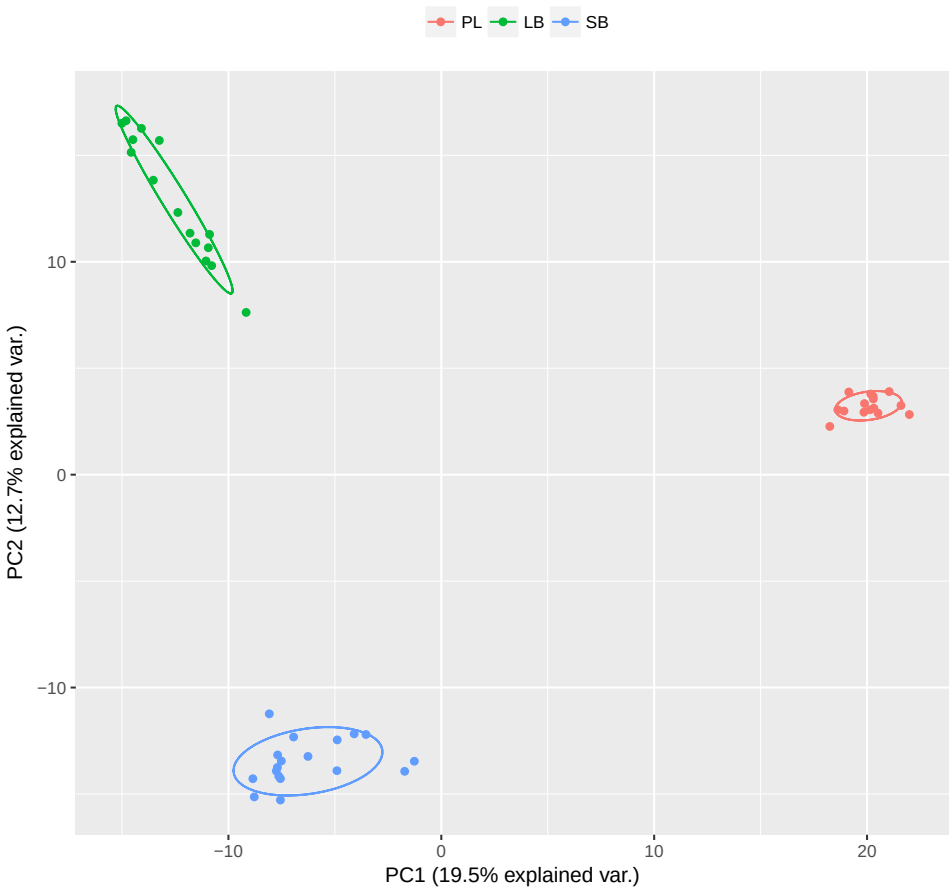


Figure 2: Genetic separation of samples from three sympatric Arctic charr morphs, based on principal component analysis of 19,252 transcriptome variants. The first and second principal components are shown with the proportion of variance explained. Individual samples are graphed (colored by color, SB: blue, LB: green and PL: red), and overlaid are 68% normal data ellipses for each morph.

## 3.2 Genetic differences between benthic and limnetic morphs in collagen metabolism and environmental sensing

In order to gauge if certain biological systems differed between the morphs we tested for GO-enrichment in variants associating with specific morphs. As some variants were annotated to different transcript isoforms of the same gene, testing was done both on transcripts and gene level using annotation to salmon genes. No GO categories were significant for SB-specific and LB-specific variants. However, PL-specific variants were significantly enriched in 10 GO-categories (Table 2). Nine categories were found on the transcipts level and seven on the gene level. Only one of the categories significant on gene level was not at the transcript level (*collagen fibril organization*) but it was close to being significant ($FDR = 0.015$). Also, variants were enriched in one other category related to the extracellular matrix *collagen catabolic process*. That may have led to two higher level categories to be significant at the transcript level but not at the gene level (*multicellular organismal catabolic process* and *macromolecule metabolic process*). Interestingly, four of the categories relate to environmental sensing and responses, *i.e.* light and sound, and showed a strong signal both on transcript and gene level, *e.g. inner ear morphogenesis* and *visual perception*. The final two categories were *tooth mineralization* and *odontogenesis*, both related to tooth development.

GO-enrichment tests on private variants only revealed three GO-categories related to tRNA aminoacylation significant at the transcript level for variants private to PL-charr. The signal is most likely because three transcripts are annotated to the same gene rather than being a real biological pattern (Table S4). Together, the GO analyses

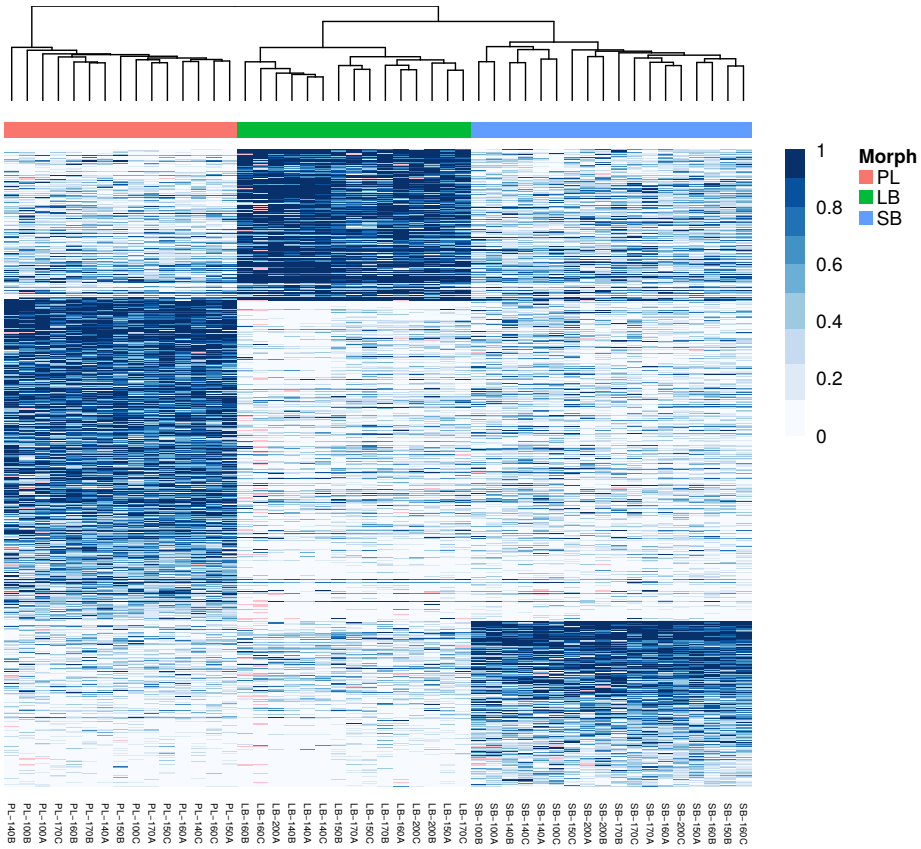


Figure 3: Separation of morphs based on the 2331 variants with $F_{ST}$ >0.2. Each column represents a sample, name below indicates morph, developmental timepoint and biological replicate. The white-blue scale depicts allele frequency, higher frequency of the alternative allele with darker blue. Hierarchical clustering grouped samples and variants. This separated the morphs (abscissa) and similar variants by allele frequency (ordinate). A total of 1174 variants had higher frequency in PL-charr, 552 in LB- and 605 in SB-charr. Missing values are indicated by pink. Coloring of individuals by morph, SB: blue, LB: green and PL: red.

of biological functions point to genetic differences in sensing, collagen metabolism and mineralization, between benthic and limnetic morphs. This could be linked to distinct differences in the main feeding habitats and principal prey species of PL- versus LB- and SB-charr and to differences in craniofacial features of these morphs.

## 3.3 Genome-wide divergence of sympatric charr morphs

To explore if the genetic differences between morphs were restricted to specific chromosomal regions or more widely distributed, we mapped the variants to the Arctic charr genome using `blastn` (see methods). A large fraction (93%, 17933 of 19,252) mapped to the genome. Of the variants mapped to the genome 10,956 or 61% mapped to chromosomes but the rest mapped to unplaced scaffolds and contigs. The number of variants found for each linkage group was, as expected, related to the linkage group size (Figure S2). Markers with $F_{ST} > 0.2$ were found on all linkage groups, and peaks of more pronounced differentiation were found on many chromosomes. Similarly, the three types of high $F_{ST}$ markers (PL-, SB- and LB-charr specific) were found on almost all linkage groups (Figure 4, B).

There were indications of enrichment of high $F_{ST}$ variants of particular type in particular linkage groups, most dramatically on the mitochondrial chromosome. Previous study of genetic differentiation in the transcriptomes of SB-charr and Aquaculture charr implicated morph differences in the mtDNA (Gudbrandsson *et al.* 2016), and

Table 2: Gene ontology categories enriched in variants with $F_{ST}$ above 0.2 and highest mean allele frequency deviation for PL-charr on the transcripts ($tr$) and gene ($ge$) level. The number of transcripts and genes with high divergence is shown ($PL_{tr}$ and $PL_{ge}$) and the total number of transcripts and genes tested in the category as well ($Tot_{tr}$ and $Tot_{ge}$). The multiple testing corrected p-value or false discovery rate ($FDR$) is also shown for both levels.

| Category | Term | $PL_{tr}$ | $Tot_{tr}$ | $FDR_{tr}$ | $PL_{ge}$ | $Tot_{ge}$ | $FDR_{ge}$ |
|---|---|---|---|---|---|---|---|
| GO:0007601 | visual perception | 30 | 122 | 0.0020 | 32 | 114 | 0.0014 |
| GO:0050953 | sensory perception of light stimulus | 30 | 122 | 0.0020 | 32 | 114 | 0.0014 |
| GO:0007605 | sensory perception of sound | 34 | 156 | 0.0025 | 36 | 137 | 0.0014 |
| GO:0034505 | tooth mineralization | 10 | 18 | 0.0025 | 9 | 13 | 0.0031 |
| GO:0030574 | collagen catabolic process | 23 | 89 | 0.0069 | 20 | 60 | 0.0071 |
| GO:0042472 | inner ear morphogenesis | 30 | 137 | 0.0083 | 32 | 128 | 0.0087 |
| GO:0030199 | collagen fibril organization | 19 | 73 | 0.0150 | 18 | 51 | 0.0073 |
| GO:0044243 | multicellular organismal catabolic process | 23 | 93 | 0.0087 | 20 | 65 | 0.0140 |
| GO:0042476 | odontogenesis | 26 | 109 | 0.0083 | 24 | 92 | 0.0359 |
| GO:0043170 | macromolecule metabolic process | 499 | 4,790 | 0.0083 | 490 | 4,155 | 0.3650 |

genotyping of population samples revealed SNPs with high frequency derived alleles in LB- and SB-charr. In this transcriptome, we found 6 variants within the mtDNA (Table S1), and all but one had high $F_{ST}$ (Figure 4). Consistent with previous results, these variants were LB- or SB-specific. Three markers (m1829G>A, m3211T>C and m3411C>T) within the 12s and 16s rRNA gene were observed previously (Gudbrandsson *et al.* 2016). Smaller fraction of markers had high $F_{ST}$'s on other linkage groups. The distribution of markers indicated distinct chromsomal regions with high differentiation associating with particular morphs. For instance a high peak of PL-specific variants was on LG18, and a peak of SB-specific markers on LG10 (Figure 4). On LG1 high $F_{ST}$ variants of all categories (PL, SB, LB specific) were found. In sum, the genetic separation between morphs was found on all linkage groups, including strong differences in the mitochondrial DNA.

### 3.4 Verification of variants in population sample confirms morph separation

As the candidate variants were derived from sequenced pools of embryos from multi-parent families, we wanted to verify them in population samples from the wild to address three questions: First, do estimates of allele frequencies in the transcriptome and in wild populations correspond? Second, do the three morphs differ genetically in population samples? Third, does the understudied Piscivorous (PI) charr differ genetically from the other three morphs or could it be an ontogenetic morph of PL-charr that learn to utilize threespine sticklebacks as prey (Snorrason *et al.* 1989)?

Candidate variants (23 in total, Table S5) were chosen based on high $F_{ST}$ values in the transcriptome sequencing, chromosomal location, biological functions and/or differential expression (Guðbrandsson *et al.* 2018). Sexually mature (some running) individuals of all four morphs in Lake Þingvallavatn were genotyped. All markers but one amplified successfully and behaved as true single nucleotide polymorphisms (SNPs, Table S5). The allele frequencies in the transcriptome and population sample were highly correlated (Kendall's $\tau = 0.71$, $p < 0.0001$) (Figure S5A). The same applied to $F_{ST}$ values (Kendall's $\tau = 0.60$,$p < 0.001$, Figure S5B). Allele frequencies seemed to be underestimated in the transcriptome, particularly lower frequencies (< 0.3) (Figure S5A). The overall $F_{ST}$'s for individual markers ranged from 0.01 (*gnl3l_G795T*) to 0.66 (*gas1l_A3641C* and *wee1_T305A*) (Table S7). The correspondence of allele frequencies in the transcriptome and the population sample, suggests patterns in the former are real. Majority of markers (18/22) deviated significantly from Hardy-Weinberg equilibrium when tested on the 93 samples (all four morphs), but almost all were in HWE within each morph (Table S7). Furthermore, majority of the variance in genotype frequency associated with morph (Figure 5, first two principal components (PCs) explain 59.6% of the variance). PC1 (45.2% of variance explained) separated benthic-limnetic morphs while PC2 (14.4%) distinguished the three transcriptome morphs (SB, LB and PL-charr). The HWE tests and PC analyses both indicated genetic differentiation of the three morphs. Note, the higher proportion of variance explained by the first two PC's for the genotyped SNPs (compared to transcriptome) likely reflects the non-random choice of variants with large frequency differences for genotyping.

The phyletic relationship of the rare PI-charr to the other morphs is unknown (Gíslason 1998; Volpe & Ferguson 1996). No transcriptome data were generated for PI-charr, but 21 individuals were genotyped. Contrary to

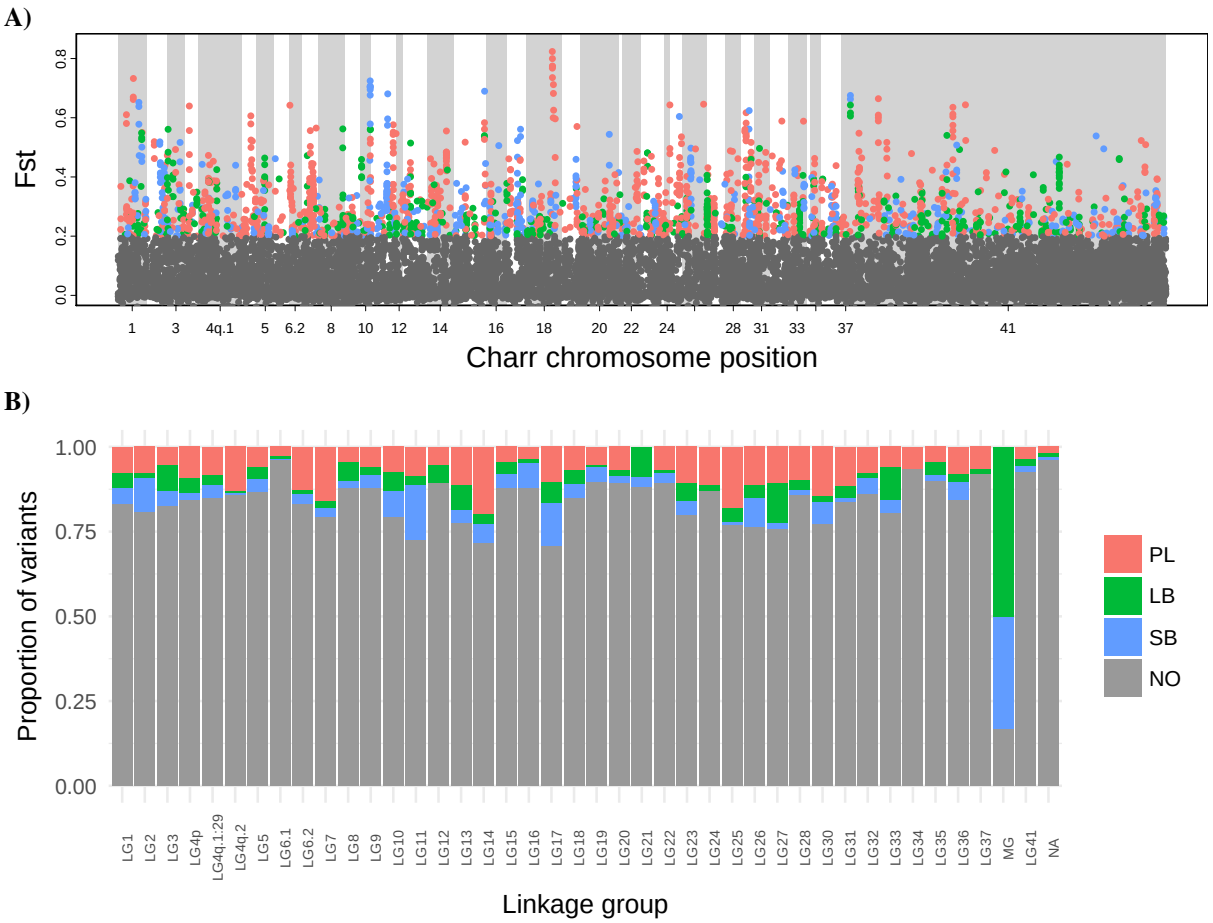


Figure 4: **A)** $F_{ST}$ values plotted by position of variants on the charr genome from (Christensen *et al.* 2018). The colors indicate which morph differs most strongly in allele frequency from the other two for variants with $F_{ST}$ above 0.2. Red PI, blue SB, green LB and gray represents variants with $F_{ST}$ below 0.2. **B)** Proportion of each variant group on each linkage group (MG: mitochondrial chromosome). Unplaced scaffolds and contigs are represented by "linkage" group 41 and in **B)** NA refers to unmapped markers.

the other three morphs, PI individuals did not group in the PC plot, suggesting genetic heterogeneity. Half of the PI-charr were in the PL-cluster but the rest (except one) fell inbetween the PL- and LB-clusters. Thus while the other three morphs separated clearly genetically, the genetic status and nature of the piscivorous charr remains in doubt.

## 3.5 Linkage disequilibrium in the population samples and peaks of differentiation

Finally, we asked how linkage disequilibrium (LD) between markers in the populations samples associated with peaks of differentiation. We analyzed variants from the transcriptome and population samples. Variants were chosen for genotyping to i) tag distinct linkage groups and ii) study segregation of linked variants on specific linkage groups.

The genotyped variants mapped to 12 separate linkage groups and 3 unplaced scaffolds (Supplemental table S5). We picked eleven markers that separate benthic-limnetic morphs (Figure 6), but despite strong associations, no marker coupled fully to a morph or morphotype. Several putatively linked markers segregated together, for instance three markers mapping to LG18. All separated PL-charr from the benthic morphs (Figure 7 A), and while *lrrc1* and *cox11* are 2.8 Mb apart the markers showed tight LD ($r^2 > 0.72$) (Table 3). The third marker (*adk_C1075T*), located 3.4 Mb downstream of *cox11*, had weak or no LD with the other two (Table 3), indicating two differentiated regions. Note, these LD-estimates are imprecise because only 21 or 24 fish were sampled for each morph (Yan *et al.* 2009).

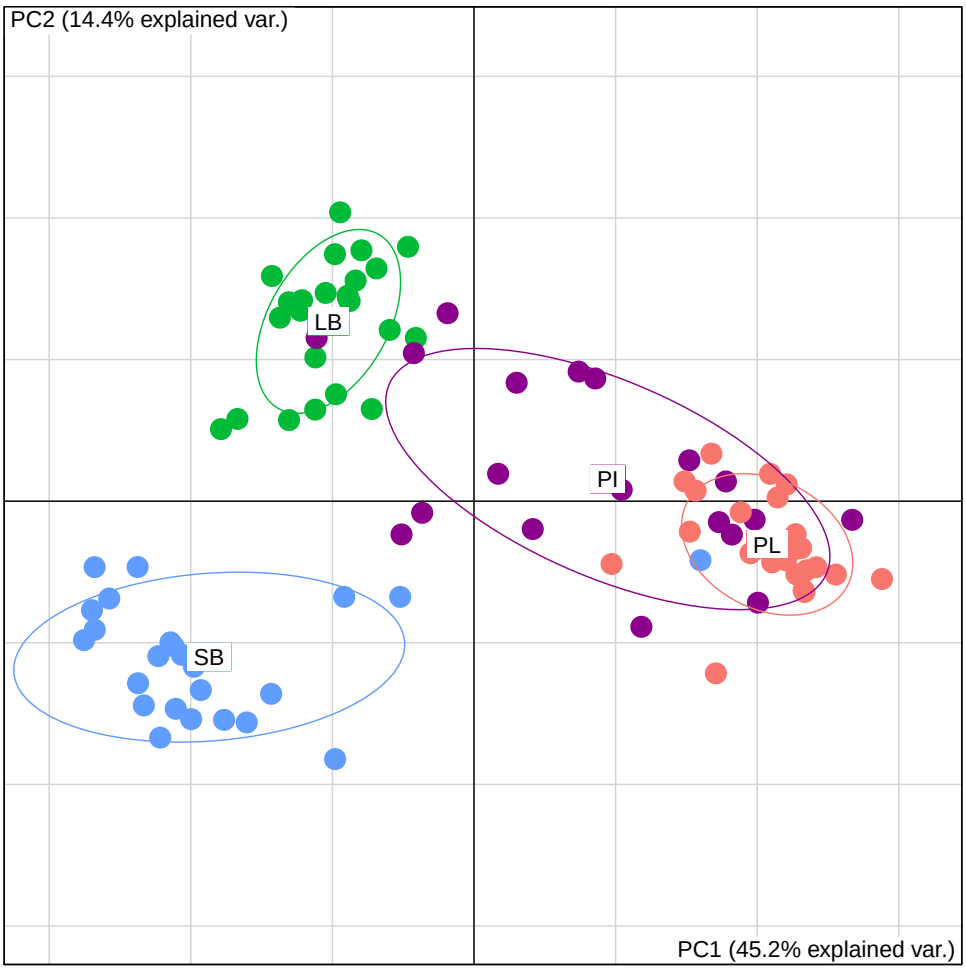


Figure 5: Genetic separation of the four morphs sympatric charr morphs depicted with principal component analyses, based on KASP genotyping of population samples. Individuals are graphed according to scores of the first two PC's on genotypes from 22 markers, along with 68% normal data ellipses for each morph (SB: blue, LB: green, PL: red and PI: purple).The PL, LB and SB-charr form distinct clusters, except one SB groups with PL.

A pair of PL-specific markers (*Kiaa1324_TC393AA* and *eif4g2b_G652A*) mapped to LG26 (Figure 7 B). As the markers are 16.8 Mb apart, very low or no association was observed between them (Table 3). The same holds for two markers in (*msi1* and *tcf15*) on LG11 that are 11.9 Mb apart and had low LD (Figure 7 C, Table 3). The region around *calm1* had numerous variants with high differentiation between PL and the two benthic morphs (Figure 7 D), but it is uncertain whether that reflects a single or many differentiated regions. Finally, three SB-specific variants (*gas1l_A3641C*, *wee1_T305A* and *dennd5a_A2555T*) were in strong LD ($r^2 > 0.45$) in all morphs, indicating long haplotypes. The variants in *dennd5a* and *wee1* mapped 118 kb apart on LG10, but on the other hand the variant in *gas1l* had the best blast hit on LG15 (The second best hit was on LG10 close to the other two, Figure 7 E). Notably, the *gas1l* variant was in linkage equilibrium with another variant on LG15 (in *tmem9b*, Figure 7 F, Table 3). The three SB-specific variants all mapped to the same chromosome in *S. salar* (ssa26) and *O. mykiss* (omy06) genome. LG10 and LG15 are ohnologous and still retain over 90% sequence similarity. This discrepancy may be due to assembly errors in the Arctic charr draft genome by Christensen *et al.* (2018), as homologous region can be tricky to assemble. Other pairs of markers mapping to different linkage groups had low LD (Table S8), the curious exception being (*cdk2ap1_A257G* and *gnl3l_G795T*) which had strong LD ($r^2$ > 0.85) in three morphs but not PL-charr ($r^2$ = 0.09) (Table 3). These two markers map to non-homologous linkage groups (LG20 and LG1). The reason for these apparent differences in LD between morphs could be chance, sampling effects, or differences in chromosomal structure between morphs. The validated variants were often in close proximity to other variants not taken for validation with strong allele differences between morphs (Figure S6). In summary, the results suggest no single variant associated fully with specific morph, but indicate that many

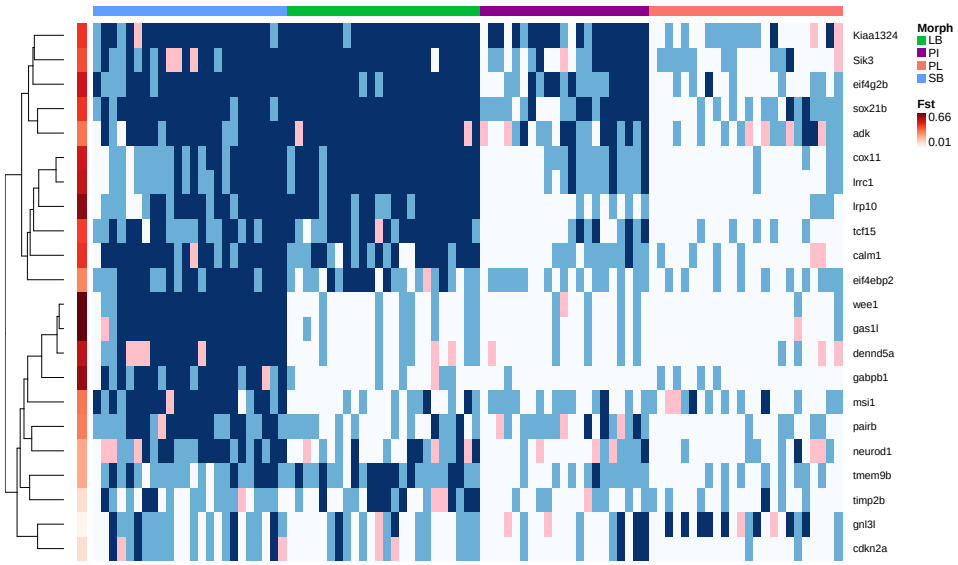


Figure 6: Heatmap of genetic variation in population samples of four charr morphs, depicting association of variants with morphology and linkage of markers. Genotypes of 22 loci were clustered by genes. $F_{ST}$ values (shades of red) for morphs are graphed for each marker. The morphs are color coded, SB: blue, LB: green, PL: red and PI: purple, and genotypes, homozygous reference allele (white), heterozygous (light-blue) and homozygous alternate allele (blue), with pink indicating missing data.

distinct chromosomal regions harbor variants separating the morphs.

# 4 Discussion

The four sympatric morphs of Arctic charr in Lake Þingvallavatn differ in a host of phenotypic attributes, e.g. in adult morphology, habitat choice, feeding preferences, growth pattern, size and age at maturity (see references in Sandlund *et al.* 1992). Common garden rearing experiments indicate these differences are based on genotype but also influenced by the environment, and emerge during embryonic and juvenile development (Skúlason *et al.* 1989a; Parsons *et al.* 2010; Kapralova *et al.* 2015) While population genetics indicated genetic separation of PL-, LB- and SB-charr (Magnusson & Ferguson 1987; Volpe & Ferguson 1996), the degree of relatedness and phyletic relationships of these morphs have remained unresolved. Here we utilized SNPs from RNAseq data derived from embryos of pure crosses of three morphs (Guðbrandsson *et al.* 2018) to assess their genetic separation. For each variant we asked if one of the morphs was more diverged, and if so, how morph specific variants were distributed in the genome and whether they related to particular functional systems. Although the experimental design, pools of individuals from bulk crosses of each morph, was not optimal for these purposes, and despite the fact that the transcriptome suffered from 3'bias due to mRNA degradation (Guðbrandsson *et al.* 2018), the data yielded many informative variants. Consistently, 22 of the 23 selected variants genotyped in independent population samples were real variants. According to the data, estimates of $F_{ST}$ in the transcriptome may have been inflated, relative to the population samples (16 out of 22 markers had lower $F_{ST}$ in the latter). Notably, the frequencies of rare alleles were underestimated in the transcriptome as expected (Konczal *et al.* 2014), but might be exaggerated due to the sequencing of related individuals, pooling of three embryos in individual samples, or Beavis effects (Beavis 1994) because high $F_{ST}$ SNPs were chosen for verification. These facts limit the interpretation of this dataset, precluding for example estimation of the site frequency spectrum, but do not negate the observed widespread genetic differentiation between morphs. Previously, we (Gudbrandsson *et al.* 2016) extracted variants from RNAseq data from SB-charr of Þingvallavatn and an aquaculture breeding stock derived from several populations (mainly from the north of Iceland (Svavarsson 2007) - an outgroup for Þingvallavatn charr). While that study yielded a similar number of candidate variants ($\sim$ 20,000) as were found here, the proportion of variants with high $F_{ST}$ was larger, probably reflecting more divergence between the aquculture charr and the SB-charr than among the three Þingvallavatn morphs studied here (and consistent with monophyletic status

Table 3: Estimates of LD ($r^2$), within each morph, and genomic distance for pairs of genotyped variants that mapped to the same linkage group or showed strong association.

| Var1 | Var2 | Pattern of divergence[a] | Linkage group | Genomic distance (bp) | LB | SB | PL | PI |
|---|---|---|---|---|---|---|---|---|
| *lrrc1* | *cox11* | PL | LG18 | 2,762,681 | 1.00 | 0.91 | 0.73 | 0.90 |
| *cox11* | *adk* | PL | LG18 | 3,370,151 | NA | 0.12 | 0.05 | 0.23 |
| *lrrc1* | *adk* | PL | LG18 | 6,132,832 | NA | 0.10 | 0.00 | 0.32 |
| *eif4g2b* | *Kiaa1324* | PL | LG26 | 16,793,974 | 0.00 | 0.03 | 0.01 | 0.06 |
| *msi1* | *tcf15* | SB, PL | LG11 | 11,914,739 | 0.06 | 0.01 | 0.04 | 0.00 |
| *dennd5a* | *wee1* | SB | LG10 | 118,070 | 0.92 | 0.82 | 0.53 | 1.00 |
| *wee1* | *gas1l* | SB | LG10[b] | [b]210,891 | 0.85 | 0.77 | 0.51 | 0.95 |
| *dennd5a* | *gas1l* | SB | LG10[b] | [b]328,961 | 0.79 | 0.63 | 0.46 | 0.95 |
| *gas1l* | *tmem9b* | SB, PL | LG15 | 400,866 | 0.00 | 0.01 | 0.00 | 0.26 |
| *cdk2ap1*[c] | *gnl3l* | SB, SB | LG20 / LG1[c] | | 0.85 | 0.97 | 0.09 | 0.88 |

[a] In which morph did the variant differ most in frequency, first description for Var1 and second for Var2.
[b] Second best blast hit for *gas1l* was on LG10, the best was on LG15.
[c] The LD for this pair was reported because of the strong association, though the markers do not map to the same chromosome.
NA: Not available, because variant *adk* was not polymorphic in LB.

of the Lake Þingvallavatn morphs compared to other Icelandic populations (Kapralova *et al.* 2011)). It remains to to be determined how many of the variants observed in these RNAseq data are shared with other populations in Iceland and across the species range (Brunner *et al.* 2001).

## 4.1 Genetic separation of recently evolved sympatric morphs

Geological forces, volcanic activity, isostatic rebound as well as underground rivers and springs in the rift-zone generate specific niches for the Icelandic biota (Kornobis *et al.* 2010; Kristjánsson *et al.* 2012; Marteinsson *et al.* 2013; Gudmundsdóttir *et al.* 2017), but also perturb them. Lake Þingvallavatn resides in a geologically active rift zone with the most recent eruption approx. 2000 years ago (Saemundsson 1992). Furthermore, the tempo of the Ice-age glacial retreat was uneven, interleaved with periods of glacial advance (Saemundsson 1992; Norðdahl *et al.* 2008). Thus, the history of colonization and adaptations by Arctic charr and other organisms in this geographic region may be quite complex.

Studying the relatedness of the morphs can shed some light on the history of colonization and divergence in Lake Þingvallavatn. The morphs (PL-, LB- and SB-charr) are more closely related to one another than to other Icelandic populations (Kapralova *et al.* 2011). The data presented here firmly reject hypothesis I, that the PL-, LB- and SB-charr morphs are formed by environmental influence on individuals of a shared gene pool (the PI-charr may be an exception, see below). The data also suggest that the benthic morphs (LB- and SB-charr) are more closely related. This is congruent with patterns of differential expression in this transcriptome (Guðbrandsson *et al.* 2018) and one previous study (Volpe & Ferguson 1996) but not others (Magnusson & Ferguson 1987; Danzmann *et al.* 1991; Gíslason 1998). This incongruence and the observed distribution of the high $F_{ST}$ variants into PL-, LB- and SB-specific categories, may reflect limited resolution due to few markers in earlier studies, incomplete sorting of alleles or differential effects of selection on distinct traits and variants. Coalescence simulations based on microsatellite data were more consistent with a brief phase of allopatric separation of the ancestors of PL- and SB-charr, rather than sympatric evolution (Kapralova *et al.* 2011). This study considered two demographic scenarios and two parameters (migration and population size). Analyses of other scenarios, *i.e.* with changing migration rates or introgression would be most interesting (see for example Jacobs *et al.* 2018). Considering the episodic nature of ice retreat and the high geological activity in the Þingvallavatn area during the formation of the lake (Saemundsson 1992; Norðdahl *et al.* 2008) charr may have invaded the waters more than once, or experienced intromittent isolation of populations in rifts. The current study lacks a natural outgroup, and future population genomic surveys of these morphs should include populations from the neighboring geographic area, including the Þingvallavatn outflow river Sog and its downstream Lake Úlfljótsvatn. That could provide better estimates of relatedness, gene flow and history of these populations and morphs. Lake Úlfljótsvatn is particularly interesting as it also hosts landlocked morphs of charr, including forms similar to the PL- and SB-charr (Jóhannsson *et al.* 1994;

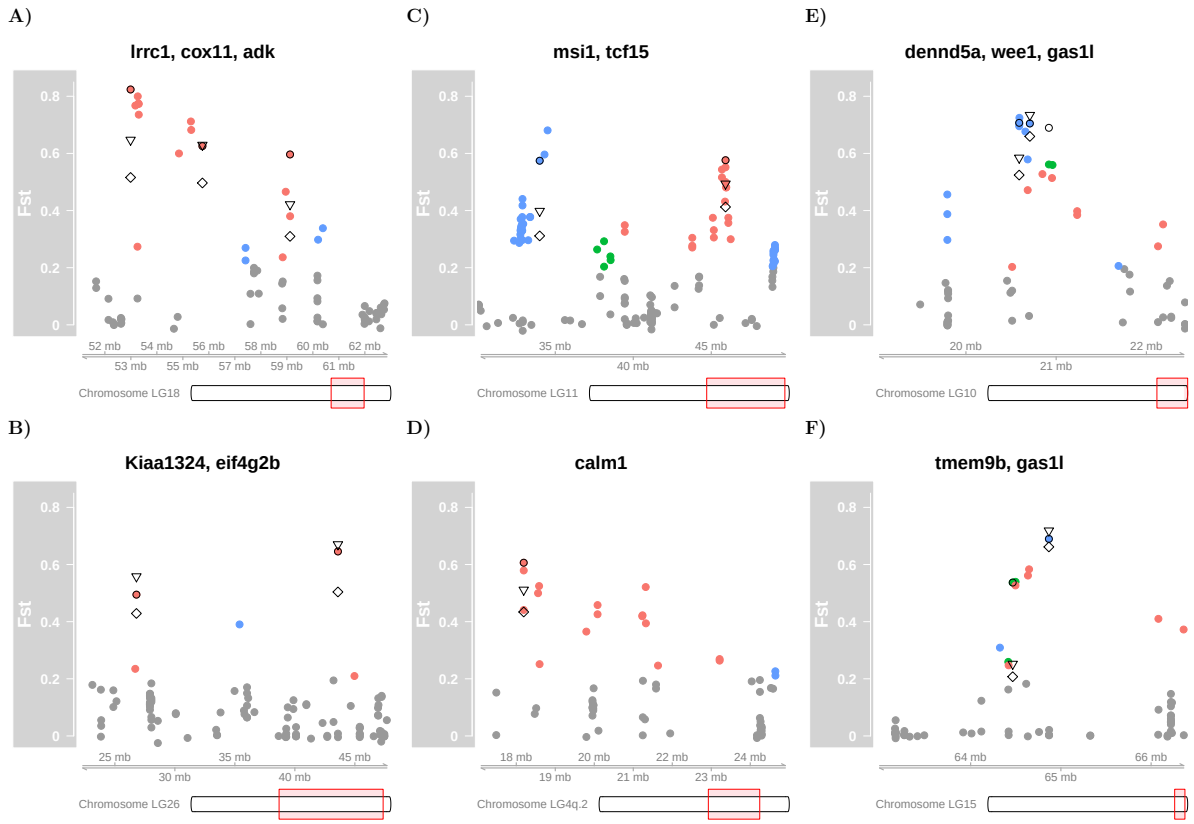


Figure 7: Detailed view of chromosomal regions and variants showing high divergence in the KASP-assay. **A)** A region on LG18 that includes *lrcc1*, *cox11* and *adk*. **B)** A large region on LG26 that includes *Kiaa1324* and *eif4g2b*. **C)** A region on LG11 that includes *msi1* and *tcf15*.**D)** Region on that includes *calm1* and many other variants with high $F_{ST}$ values that were not validated. **E)** A region on LG10 that includes *dennd5a*, *wee1*. The $F_{ST}$-value from the transcriptome for the variant in *gas1* is shown at the location of the second best blast hit with a black ring. **F)** A region on LG15 that includes *tmem9b* and *gas1l*. The regions on LG10 and LG15 shown in **E)** and **F)** are homologous and some of the markers might belong better in the opposite linkage group as we suggest is the case for *gas1l*. The colored dots mark the values from the transcriptome as in Fig. 4 and the variants taken for validation are marked with a black ring (○). The triangles (▽) show the $F_{ST}$ value from the KASP-assay for the three transcriptome morphs (PL, SB and LB) and the diamonds (◇) $F_{ST}$ for all morphs (including PI).

Woods *et al.* 2012a).

The *Salvelinus* genus is particularly interesting for sympatric polymorphism (Noakes 2008; Klemetsen 2010; Muir *et al.* 2016; Markevich *et al.* 2018) that seems to run counter to the predominant mode of allopatric speciation. Smith & Skúlason (1996) argued resource polymorphism and phenotypic plasticity could potentiate ecological speciation, even within a geographic area. The question remains if charr morphs in sympatry (e.g. in the Transbaikalian lakes (Gordeeva *et al.* 2015), Lake Galtarból in northern Iceland (Wilson *et al.* 2004), and Lower Tazimina Lake in Alaska (May-Mcnally *et al.* 2015)) differentiated in true sympatry, or originate by repeated invasions of the waterbodies. The morphs in Loch Stack (Adams *et al.* 2008) and Loch Tay (Garduño-Paz *et al.* 2012) are clearly of allopatric origin. A recent genome-wide study of sympatric charr in Scotland and Siberia is consistent with sympatric origins of charr morphs (Jacobs *et al.* 2018). Ecological specialization can lead to differential growth, behavior and maturity that may influence both choice of spawning sites and timing of spawning. Within Lake Þingvallavatn, the timing of spawning differs between the three best studied morphs. The LB-charr spawns in August, PL-charr in October, but SB-charr over a broader period from September to November (Skúlason *et al.* 1989b; Sandlund *et al.* 1992). Although all morphs appear to spawn in the same habitat, i.e. among loose stones in the littoral zone, there may be micro-habitat preferences for spawning sites. Variation in mating behavior can in principle lead parents to choose specific nesting sites, e.g. close to inflow of cold water. Such differences in environmental conditions could, in turn, influence development of ecological functional traits and adult form of

the morphs.

The ambiguous genetic status of the large, rare and piscivorous morph is quite intriguing. PI-charr was only studied by genotyping 22 markers, but curiously PI individuals did not cluster as tightly as the other morphs. PI individuals grouped with LB, PL-charr and a few in between. Previous studies found PI individuals affiliated with either or both morphs (Magnusson & Ferguson 1987; Volpe & Ferguson 1996). Several models can account for this pattern. First, the classification of the morph based on phenotype may not be stringent enough and thus some LB-charr might have been misidentified as PI-charr. As the LB- and PI-charr utilize different prey and differ in spawning periods (the former in August and latter in October (Skúlason *et al.* 1989b)) this misidentification is unlikely. However, it is possible that some LB-charr males might be erroneously classified as PI because of an extended jaw-hook, if they are still running in October (some are). Secondly, the PI-charr may be more heterogeneous at the genetic level. Snorrason *et al.* (1989) postulated that PI-charr emerge as sizable PL-charr learn to eat fish. PI-charr would then be an environmentally induced morph, genetically identical to PL-charr. Thirdly, it is possible that PI-charr are genetically distinct and have either recently evolved or cross hybridize with other morphs. In this last scenario, there might be directional gene-flow from, for instance from LB-charr into PI-charr. Currently, we can not distinguish between these possibilities.

We conclude that three of the Lake Þingvallavatn morphs differ genetically, and considering the differences in trophic traits and spawning, may even be reproductively isolated (hypothesis III). It remains to be determined if the PI-charr are genetically separated from the other morphs or if they are an environmentally induced form as may be the case of eco-morphs in Sockeye salmon *Oncorhynchus nerka* (Limborg *et al.* 2018).

## 4.2 Genome-wide separation of sympatric morphs

How extensive is the genetic separation of morphs and how is the genetic differentiation distributed in the genome? For instance, are differentiating variants localized to few islands (Nadeau *et al.* 2012; Andrew & Rieseberg 2013; Malinsky *et al.* 2015) as stated in hypothesis II or are there multiple signals on many chromosomes (Hohenlohe *et al.* 2010; Jones *et al.* 2012) according hypothesis III?

Variants with relatively high $F_{ST}$ (>0.2) mapped to all 40 *S. alpinus* linkage groups (including the mitochondrial chromosome), which supports hypothesis III. The clustering of variants, some with $F_{ST}$ >0.5, suggests specific genes/regions associate with the specializations of particular morphs. Analyses of LD of genotyped variants imply that peaks on the same linkage group are independent. The distribution of variants on linkage groups is rather even, with the exception of the mtDNA which had high fraction of LB- and SB-specific variants (consistent with earlier findings (Gudbrandsson *et al.* 2016)). This might reflect evolution of mitochondrial functions in the benthic morphs. The results suggest that the extensive genetic separation of these three morphs involves many differentiating genes on many (or all) linkage groups. While the data seem to refute hypothesis II, that few genomic islands differentiate the morphs, it is plausible that ecological specialization has driven differentiation in certain genomic regions.

Previous studies of salmonids have found genome-wide differences between populations and ecological morphs. Significant genetic differences were found between migratory and non-migratory Rainbow trout (Hale *et al.* 2013), the lake, river and stream ecotypes of sockeye salmon (Larson *et al.* 2017) and ecologically different subpopulations of salmon (Vincent *et al.* 2013; Cauwelier *et al.* 2017). Sympatric ecotypes are found in several salmonid species, e.g. whitefish (Gagnaire *et al.* 2013), but are most pronounced in Arctic charr and Lake charr *Salvelinus namaycush*. Genomic studies of the latter revealed varying degrees of genetic separation of sympatric morphs in large lakes in North America (Harris *et al.* 2015; Perreault-Payette *et al.* 2017). A study on five pairs of benthic-limnetic whitefish morphs (Gagnaire *et al.* 2013) found a correlation between genetic and phenotypic differentiation, and considerable overlap of genomic regions that differentiated morph pairs in each lake. Notably the significant degree of genetic parallelism broke up at the finer level, with different haplotypes associating with morphotype - suggesting genetic and/or allelic heterogeneity of the causative loci (Gagnaire *et al.* 2013). Recently, genome wide analyses of Scottish and Siberian Arctic charr suggested limited genetic parallelism in benthic-limnetic specializations (Jacobs *et al.* 2018). Fine scale analyses of peaks of differentiation could not be conducted with the current data, but the genome wide distribution of differentiated variants suggests restricted gene flow between at least three of the morphs (PL-, LB- and SB-charr) and that they may be reproductively isolated.

Genomic differentiation reflects the history and evolution of groups, degree and age of separation of groups, extent of gene flow and past hybridization events (Seehausen *et al.* 2014; Shapiro *et al.* 2016). But intrinsic factors, such as the nature of the genes and their biological actions, also shape the rates of differentiation and divergence, and genomic features, such as centromeres, gene density, recombination and GC content influence estimates of

sequence divergence and allele frequency differences (Seehausen *et al.* 2014; Burri *et al.* 2015; Vijay *et al.* 2017). More population genomic data and analyses are needed to disentangle the role of positive selection and intrinsic factors on patterns of differentiation in this system. More broadly, the confirmed synteny of large genomic regions (Nugent *et al.* 2017), the range of species, subspecies and ecologically distinct populations (Klemetsen 2013; Jacobs *et al.* 2018) sets the stage for future studies of the genomic and ecological correlates of divergence and polymorphism in salmonids.

## 4.3 Potential mechanisms of phenotypic and developmental differences between sympatric morphs

Which genes influence differences in size, head shape, feeding apparatus, coloration, dietary preferences, parasite load, etc. between the morphs? Genetic variation in the transcriptome tags large fraction of the potentially functional regions of the genome. The genome-wide differentiation between morphs implies polygenic basis of their differences, but major genes influencing specific traits may be segregating. The GO-analyses of variants with high $F_{ST}$s pointed to enrichment of several categories between PL-charr and the benthic morphs. This further supports polygenic roots of morph differentiation. The genomic resolution of the current data is low, and while we can describe frequency differences in variants in specific genes, it is more likely that other linked polymorphisms/loci are actually contributing to phenotypic differences.

The developmental roots of the sympatric polymorphism are unknown, but the current data can inform future studies by bringing attention to particular chromosomal regions, genes and systems. The GO results suggest the limnetic and benthic morphs differ genetically in three systems (collagen metabolism, tooth mineralization and sensory functions). The enrichment of variants in collagen organzation/catabolism and extracellular matrix categories is consistent with observed differential expression of ECM related and cartilage remodeling genes in benthic and limnetic morphs (Ahi *et al.* 2014, 2015). Curiously, one genomic region with strong differentiation contained a variant in *calm1*. Calmodulins are broadly expressed in vertebrate tissues, including bone and articular cartilage in humans (Mototani *et al.* 2005). Notably, a microarray screen identified higher expression of *Calm1* in beak primordia of developing finches with longer beaks (Abzhanov *et al.* 2006). Perhaps most intruiging, three SB-specific variants on linkage group 10 (or 15) had high $F_{ST}$'s. This might represent long haplotypes with strong differentiation between morphs, and that the small benthic phenotype may be influenced by allelic variation in this region. Curiously one gene, *Wee1* is a key regulator of the timing of mitosis and cell size, first identified in fission yeast (Nurse & Thuriaux 1980) and another *gas1* (*Growth arrest-specific protein 1*) associates with embryonic cranial skeleton morphogenesis and palate development (Seppala *et al.* 2007). Furthermore, in light of altered expression of *eif4ebp1* in SB-charr, both in muscles of adult fish (Macqueen *et al.* 2011) and developing embryos (this trancriptome (Guðbrandsson *et al.* 2018)) it is curious that the frequency of a variant in this gene differed between morphs ($F_{ST}$ = 0.28). *eif4ebp1* is involved in the insulin receptor signaling pathway and mTOR signalling (Bidinosti *et al.* 2010; Gkogkas *et al.* 2012; Banko *et al.* 2007). However, here this variant associated with benthic morphotype not small benthic charr, which is at odds with the postulation that reduced growth of the SB populations is mediated through changes in the mTOR activity. We stress the general caveat, that while markers with strong frequency differences between morphs may reside in the causative genes, its more likely that they are linked to other causative variants. Few traits have been mapped to a gene in salmonids. The exception being *vgll3* which associates with age of maturity in wild populations, with very curious sex dependent dominance of alleles (Barson *et al.* 2015; Ayllon *et al.* 2015). Unfortunately the *vgll3* gene was not transcribed in our data, but SB-specific variants with $F_{ST}$ around 0.4 are found $\sim$1 Mb from the location of *vgll3* on LG2 (Figure S6 K). Further work is needed to check if variants in *vgll3*, *tulp4* (see Larson *et al.* 2017) or the genes highlighted here contribute to phenotypic differences in charr populations. QTL and/or fine scale association mapping (Zimmerman *et al.* 2000; Palsson & Gibson 2004; Dworkin *et al.* 2005) are needed to find variants that associate with body size, growth rate and trophic traits in these sympatric charr.

## 4.4 Conclusion and future perspective

We mined genetic variation of the sections of the genome more related to function, from three of the four morphs in Lake Þingvallavatn to address questions about genetic separation, genome-wide differentiation of morphs and potential functionality of loci separating them. We formulated three hypothesis about the putative causes of morph differences: Hypothesis I stated that the morphs are environmentally induced. According to hypothesis II, ecological specialization led to few genomic island of strong genetic differentiation between morphs, with low background separation across the genome. Hypotheses III postulated substantial differentiation across the genome

due to reduced gene flow, with bulk of the genome showing separation between morphs. Estimates of genetic differentiation between the Þingvallavatn morphs have yield different results from low (Volpe & Ferguson 1996; Kapralova *et al.* 2011) to extremely high (Kapralova *et al.* 2013) depending on the marker used. Based on the genetic difference observed in this study, we conclude that three of the morphs (SB, LB and PL-charr) are distinct populations and find it highly unlikely that these morphs are environmentally induced (rejecting hypothesis I). Furthermore, we interpret the data as suggesting genome-wide differentiation over weak differentiation with few islands of differentiation (hypothesis III over II). This is supported by the facts that the background $F_{ST}$ in the transcriptome was rather high and high $F_{ST}$ peaks were found on all linkage groups (and even multiple peaks on some). The fact that spawning times of some morphs do not overlap (LB-charr spawn in August, but PL- and SB-charr in September - October) also argues that gene flow has been reduced between these morphs. Note however that hybrids of specific morphs, SB-, PL- and PI-charr can be generated in the laboratory (Kapralova 2014, Kalina H. Kapralova, Sigurdur S. Snorrason, Zophonías O. Jónsson, Arnar Pálsson *et al.* unpublished data), but their fitness and how common they are in nature is unknown. Thus we conclude that the three studied morphs in Lake Þingvallavatn are not one panmictic population and that gene flow between them has been limited. The nature of the PI-charr is still in doubt, it may have arisen by ontogenetic shift as suggested by Snorrason *et al.* (1989) or perhaps recurrent hybridizations of other morphs. Population ddRAD-seq data of sexually mature/spawning charr of all morphs confirm the clear demarcation of three of the morphs, and offer insights into the nature of the PI-charr (Han Xiao, Benjamín Sigurgeirsson, et al. unpublished data). Future studies of the fertilization success of hybrids and pure morphs, analyses of the development and fitness of hybrids and pure morph, and behavioral studies of spawning behavior, reed locations and properties, and mate choice have can cast light on potential pre- and postzygotic barriers to gene flow between morphs. In sum, the observed phenotypic and genetic differences between sympatric and locally adapted Arctic charr populations, can aid futures studies of ecological adaptation and the synthesis of evolutionary, ecological and developmental biology.

# 5 Acknowledgments

Special thanks to present and past members of the Arctic charr group at the University of Iceland for support, comments and suggestions. Thanks to Ben Sutherland and Eric Normandeau for excellent advice and open communication about MapComp and methods for placing variants on linkage maps. This project was supported by The Icelandic Center for Research (RANNIS #100204011) to SSS and coworkers, The University of Iceland Doctoral Fund to JG and University of Iceland research fund to AP, SSS and ZOJ.

# 6 Author contributions

Conceived and designed the study: JG, AP, ZOJ, SSS, SRF, KHK
Sampling: SSS, KHK, SRF, ZOJ, AP, JG
Gathered the data: ZOJ, SRF, KHK, VH, ÞMB
Retrieval and analyses of variants data: JG
SNP confirmation: KHK, AP, VH, ÞMB, JG
Analyses: JG, AP, KHK
Writing: AP, JG, SSS, KHK, SRF, ZOJ

# References

Abouheif E, Favé MJ, Ibarrarán-Viniegra AS, Lesoway MP, Rafiqi AM, Rajakumar R (2014) Eco-Evo-Devo: The Time Has Come. In *Ecological Genomics - Ecology and the Evolution of Genes and Genomes* (edited by CR Landry, N Aubin-Horth), chap. 6, pp. 107–125. Springer Netherlands, Dordrecht. 2

Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature*, **442**, 563–567. 2, 18

Adams CE, Wilson AJ, Ferguson MM (2008) Parallel divergence of sympatric genetic and body size forms of Arctic charr, *Salvelinus alpinus*, from two Scottish lakes. *Biological Journal of the Linnean Society*, **95**, 748–757. 16

Ahi EP, Gudbrandsson J, Kapralova KH, *et al.* (2013) Validation of Reference Genes for Expression Studies during Craniofacial Development in Arctic Charr. *PloS ONE*, **8**, e66389. 3

Ahi EP, Kapralova KH, Pálsson A, *et al.* (2014) Transcriptional dynamics of a conserved gene expression network associated with craniofacial divergence in Arctic charr. *EvoDevo*, **5**, 40. 3, 18

Ahi EP, Steinhäuser SS, Pálsson A, *et al.* (2015) Differential expression of the aryl hydrocarbon receptor pathway associates with craniofacial polymorphism in sympatric Arctic charr. *EvoDevo*, **6**, 27. 3, 18

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the Evolution of Salmonid Fishes. In *Evolutionary Genetics of Fishes* (edited by BJ Turner), pp. 1–53. Springer US. 4, 6, 8

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–10. 6

Andrew RL, Rieseberg LH (2013) Divergence is focused on few genomic regions early in speciation: Incipient speciation of sunflower ecotypes. *Evolution*, **67**, 2468–2482. 17

Ayllon F, Kjærner-Semb E, Furmanek T, *et al.* (2015) The *vgll3* Locus Controls Age at Maturity in Wild and Domesticated Atlantic Salmon (*Salmo salar* L.) Males. *PLOS Genetics*, **11**, e1005628. 18

Banko JL, Merhav M, Stern E, Sonenberg N, Rosenblum K, Klann E (2007) Behavioral alterations in mice lacking the translation repressor 4E-BP2. *Neurobiology of Learning and Memory*, **87**, 248–256. 18

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT (2011) BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691–1692. 6

Barson NJ, Aykanat T, Hindar K, *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, **528**, 405–408. 18

Beavis WD (1994) The power and deceit of QTL experiments: Lessons from comparitive QTL studies. In *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference* (edited by DB Wilkinson), pp. 250–266. American Seed Trade Association, Washington DC. 14

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300. 7

Bernatchez L, Renaut S, Whiteley AR, *et al.* (2010) On the origin of species: Insights from the ecological genomics of lake whitefish. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 1783–800. 2, 3

Berthelot C, Brunet F, Chalopin D, *et al.* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature communications*, **5**, 3657. 3, 4, 6

Bidinosti M, Ran I, Sanchez-Carbente MR, *et al.* (2010) Postnatal Deamidation of 4E-BP2 in Brain Enhances Its Association with Raptor and Alters Kinetics of Excitatory Synaptic Transmission. *Molecular Cell*, **37**, 797–808. 18

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**, 525–527. 5

Brunner PC, Douglas MR, Osinov A, Wilson CC, Bernatchez L (2001) Holarctic phylogeography of Arctic charr (*Salvelinus alpinus* L.) inferred from mitochondrial DNA sequences. *Evolution*, **55**, 573–586. 15

Burri R, Nater A, Kawakami T, *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research*, **25**, 1656–1665. 18

Carlson M (2015) GO.db: A set of annotation maps describing the entire Gene Ontology. 7

Cauwelier E, Gilbey J, Sampayo J, Stradmeyer L, Middlemas SJ (2017) Identification of a single genomic region associated with seasonal river return timing in adult Scottish Atlantic salmon (*Salmo salar* L.) identified using a genome-wide association study. *Canadian Journal of Fisheries and Aquatic Sciences*, pp. cjfas–2017–0293. 17

Christensen KA, Rondeau EB, Minkley DR, *et al.* (2018) The Arctic charr (*Salvelinus alpinus*) genome and transcriptome assembly. *PLOS ONE*, **13**, e0204076. 3, 4, 7, 12, 13

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157. 2

Danzmann RG, Cairney M, Davidson WS, *et al.* (2005) A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). *Genome*, **48**, 1037–51. 4

Danzmann RG, Ferguson MM, Skulason S, Snorrason SS, Noakes DLG (1991) Mitochondrial DNA diversity among four sympatric morphs of Arctic charr, *Salvelinus alpinus* L., from Thingvallavatn, Iceland. *Journal of Fish Biology*, **39**, 649–659. 3, 15

Darwin C (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.*, vol. 5. John Murray, London. 1

De Meyer J, Christiaens J, Adriaens D (2016) Diet-induced phenotypic plasticity in European eel (*Anguilla anguilla*). *Journal of Experimental Biology*, **219**, 354–363. 2

Dworkin I, Palsson A, Gibson G (2005) Replication of an *Egfr*-wing shape association in a wild-caught cohort of *Drosophila melanogaster*. *Genetics*, **169**, 2115–2125. 18

Foote AD, Vijay N, Ávila-Arcos MC, *et al.* (2016) Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, **7**, 11693. 2

Frandsen F, Malmquist HJ, Snorrason SS (1989) Ecological parasitology of polymorphic Arctic charr, *Salvelinus alpinus* (L.), in Thingvallavatn, Iceland. *Journal of Fish Biology*, **34**, 281–297. 3

Fruciano C, Franchini P, Kovacova V, *et al.* (2016) Genetic linkage of distinct adaptive traits in sympatrically speciating crater lake cichlid fish. *Nature Communications*, **7**, 12736. 2

Gagnaire PA, Pavey SA, 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**, 2483–2497. 17

Garduño-Paz MV, Adams CE, Verspoor E, Knox D, Harrod C (2012) Convergent evolutionary processes driven by foraging opportunity in two sympatric morph pairs of Arctic charr with contrasting post-glacial origins. *Biological Journal of the Linnean Society*, **106**, 794–806. 16

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint*, **7**. 6

Gíslason D (1998) *Genetic and Morphological Variation in Polymorphic Arctic Charr* Salvelinus Alpinus*, from Icelandic Lakes*. Master's thesis, The University of Guelph. 3, 11, 15

Gkogkas CG, Khoutorsky A, Ran I, *et al.* (2012) Autism-related deficits via dysregulated eIF4E-dependent translational control. *Nature*, **493**, 371–377. 18

Gordeeva NV, Alekseyev SS, Matveev AN, Samusenok VP, Taylor E (2015) Parallel evolutionary divergence in Arctic char *Salvelinus alpinus* complex from Transbaikalia: Variation in differentiation degree and segregation of genetic diversity among sympatric forms. *Canadian Journal of Fisheries and Aquatic Sciences*, **72**, 96–115. 3, 16

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186. 7

Grabherr MG, Haas BJ, Yassour M, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, **29**, 644–52. 5

Grant B, Grant PR (2008) Fission and fusion of Darwin's finches populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2821–2829. 2

Grant PR, Grant BR (2002) Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science*, **296**, 707–711. 1

Gudbrandsson J, Ahi EP, Franzdottir SR, *et al.* (2016) The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs [version 3; referees: 2 approved, 1 approved with reservations]. *F1000Research*, **4**. 3, 4, 10, 11, 14, 17

Guðbrandsson J, Franzdóttir SR, Kristjánsson BK, *et al.* (2018) Differential gene expression during early development in recently evolved and sympatric Arctic charr morphs. *PeerJ*, **6**, e4345. 2, 3, 4, 5, 6, 7, 8, 11, 14, 15, 18

Gudmundsdóttir R, Kornobis E, Kristjánsson BK, Pálsson S (2017) Genetic analysis of ciliates living on the groundwater amphipod *Crangonyx islandicus* (Amphipoda: Crangonyctidae). *Acta Zoologica*, **99**, 188–198. 15

Guerrero RF, Hahn MW (2017) Speciation as a sieve for ancestral polymorphism. *Molecular Ecology*, pp. 1–7. 2

Haas BJ (2015) Trinotate: Transcriptome Functional Annotation and Analysis. 6

Haas BJ, Papanicolaou A (2015) TransDecoder (Find Coding Regions Within Transcripts). 7

Hahsler M, Nagar A (2017) rBLAST: Interface to the Basic Local Alignment Search Tool (BLAST). 7

Hale MC, Thrower FP, Berntson EA, Miller MR, Nichols KM (2013) Evaluating Adaptive Divergence Between Migratory and Nonmigratory Ecotypes of a Salmonid Fish, *Oncorhynchus mykiss*. *G3; Genes|Genomes|Genetics*, **3**, 1273–1285. 17

Harris LN, Chavarie L, Bajno R, *et al.* (2015) Evolution and origin of sympatric shallow-water morphotypes of Lake Trout, *Salvelinus namaycush*, in Canada's Great Bear Lake. *Heredity*, **114**, 94–106. 17

Hartl DL, Clark AG (2006) *Principles of Population Genetics, Fourth Edition*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA. 6

He C, Holme J, Anthony J (2014) SNP genotyping: The KASP assay. *Methods in molecular biology*, **1145**, 75–86. 7

Hendry AP, Bolnick DI, Berner D, Peichel CL (2009) Along the speciation continuum in sticklebacks. *Journal of Fish Biology*, **75**, 2000–2036. 2

Hindar K, Ryman N, Ståhl G (1986) Genetic differentiation among local populations and morphotypes of Arctic charr, *Salvelinus alpinus*. *Biological Journal of the Linnean Society*, **27**, 269–285. 3

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**. 17

Jacobs A, Carruthers M, Yurchenko A, *et al.* (2018) Convergence in form and function overcomes non-parallel evolutionary histories in Arctic charr. *bioRxiv*, p. 265272. 3, 15, 16, 17, 18

Jóhannsson M, Guðbergsson G, Antonsson T, Guðjónsson S (1994) Fiskrannsóknir á Úlfljótsvatni 1993 - VMST-S/94001X. Tech. Rep. VMST-S/94001X, Veiðimálastofnun, Selfoss. 15

Jombart T (2008) Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405. 7

Jones FC, Chan YF, Schmutz J, *et al.* (2012) A Genome-wide SNP Genotyping Array Reveals Patterns of Global and Repeated Species-Pair Divergence in Sticklebacks. *Current Biology*, **22**, 83–90. 2, 17

Jonsson B, Skúlason S, Snorrason SS, *et al.* (1988) Life History Variation of Polymorphic Arctic Charr (*Salvelinus alpinus*) in Thingvallavatn, Iceland. *Canadian Journal of Fisheries and Aquatic Sciences*, **45**, 1537–1547. 3

Kapralova KH (2008) *Evolutionary Origins of Widespread Small Benthic Charr (*Salvelinus Alpinus*) over Different Geographic Scales in Iceland*. MSc., University of Iceland, Reykjavík, Iceland. 3

Kapralova KH (2014) *Study of Morphogenesis and miRNA Expression Associated with Craniofacial Diversity in Arctic Charr (*Salvelinus Alpinus*) Morphs*. Ph.D. thesis, University of Iceland, Reykjavík, Iceland. 19

Kapralova KH, Franzdóttir SR, Jónsson H, Snorrason SS, Jónsson ZO (2014) Patterns of MiRNA Expression in Arctic Charr Development. *PLoS ONE*, **9**, e106084. 3

Kapralova KH, Gudbrandsson J, Reynisdottir S, *et al.* (2013) Differentiation at the *MHCII*$\alpha$ and *Cath2* Loci in Sympatric *Salvelinus alpinus* Resource Morphs in Lake Thingvallavatn. *PLoS ONE*, **8**, e69402. 3, 19

Kapralova KH, Jónsson ZO, Palsson A, *et al.* (2015) Bones in motion: Ontogeny of craniofacial development in sympatric arctic charr morphs. *Developmental Dynamics*, **244**, 1168–1178. 3, 14

Kapralova KH, Morrissey MB, Kristjánsson BK, Ólafsdóttir GA, Snorrason SS, Ferguson MM (2011) Evolution of adaptive diversity and genetic connectivity in Arctic charr (*Salvelinus alpinus*) in Iceland. *Heredity*, **106**, 472–87. 3, 5, 15, 19

Klemetsen A (2010) The Charr Problem Revisited: Exceptional Phenotypic Plasticity Promotes Ecological Speciation in Postglacial Lakes. *Freshwater Reviews*, **3**, 49–74. 2, 16

Klemetsen A (2013) The Most Variable Vertebrate on Earth. *Journal of Ichthyology*, **53**, 781–791. 2, 18

Konczal M, Koteja P, Stuglik MT, Radwan J, Babik W (2014) Accuracy of allele frequency estimation using pooled RNA-Seq. *Molecular Ecology Resources*, **14**, 381–392. 14

Kornobis E, Pálsson S, Kristjánsson BK, Svavarsson J (2010) Molecular evidence of the survival of subterranean amphipods (Arthropoda) during Ice Age underneath glaciers in Iceland. *Molecular ecology*, **19**, 2516–30. 15

Kristjánsson BK, Skúlason S, Snorrason SS, Noakes DLG (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**, 1099–112. 3, 15

Krueger F (2012) Trim Galore! 5

Larson WA, Limborg MT, McKinney GJ, Schindler DE, Seeb JE, Seeb LW (2017) Genomic islands of divergence linked to ecotypic variation in sockeye salmon. *Molecular Ecology*, **26**, 554–570. 17, 18

Li H, Handsaker B, Wysoker A, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079. 6

Lien S, Koop BF, Sandve SR, *et al.* (2016) The Atlantic salmon genome provides insights into rediploidization. *Nature*, **533**, 200–205. 4

Limborg MT, Larson WA, Shedd K, Seeb LW, Seeb JE (2018) Novel RAD sequence data reveal a lack of genomic divergence between dietary ecotypes in a landlocked salmonid population. *Conservation Genetics Resources*, **10**, 169–171. 17

Losos JB, Ricklefs RE (2009) Adaptation and diversification on islands. *Nature*, **457**, 830–836. 1

Lowry D, Gould B (2016) Speciation Continuum. In *Encyclopedia of Evolutionary Biology*, vol. 4, pp. 159–165. Elsevier. 2

Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20132881. 4

Macqueen DJ, Kristjánsson BK, Paxton CGM, Vieira VLA, Johnston IA (2011) The parallel evolution of dwarfism in Arctic charr is accompanied by adaptive divergence in mTOR-pathway gene expression. *Molecular ecology*, **20**, 3167–84. 3, 18

Magnusson KP, Ferguson MM (1987) Genetic analysis of four sympatric morphs of Arctic charr, *Salvelinus alpinus*, from Thingvallavatn, Iceland. *Environmental Biology of Fishes*, **20**, 67–73. 3, 14, 15, 17

Malinsky M, Challis RJ, Tyers AM, *et al.* (2015) Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, **350**, 1493–1498. 2, 17

Malmquist HJ, Snorrason SS, Skulason S, Jonsson B, Sandlund OT, Jonasson PM (1992) Diet Differentiation in Polymorphic Arctic Charr in Thingvallavatn, Iceland. *The Journal of Animal Ecology*, **61**, 21–35. 3

Markevich G, Esin E, Anisimova L (2018) Basic description and some notes on the evolution of seven sympatric morphs of Dolly Varden *Salvelinus malma* from the Lake Kronotskoe Basin. *Ecology and Evolution*, **8**, 2554–2567. 2, 16

Marteinsson VT, Rúnarsson A, Stefánsson A, *et al.* (2013) Microbial communities in the subglacial waters of the Vatnajökull ice cap, Iceland. *ISME Journal*, **7**, 427–437. 15

May-Mcnally SL, Quinn TP, Woods PJ, Taylor EB (2015) Evidence for genetic distinction among sympatric ecotypes of Arctic char (*Salvelinus alpinus*) in south-western Alaskan lakes. *Ecology of Freshwater Fish*, **24**, 562–574. 16

Moghadam HK, Ferguson MM, Danzmann RG (2011) Whole genome duplication: Challenges and considerations associated with sequence orthology assignment in Salmoninae. *Journal of Fish Biology*, **79**, 561–574. 6, 8

Morris M, Rogers SM (2014) Integrating Phenotypic Plasticity Within an Ecological Genomics Framework: Recent Insights from the Genomics, Evolution, Ecology, and Fitness of Plasticity. In *Ecological Genomics - Ecology and the Evolution of Genes and Genomes* (edited by CR Landry, N Aubin-Horth), chap. 5, pp. 73–105. Springer Netherlands, Dordrecht. 2

Mototani H, Mabuchi A, Saito S, *et al.* (2005) A functional single nucleotide polymorphism in the core promoter region of *CALM1* is associated with hip osteoarthritis in Japanese. *Human molecular genetics*, **14**, 1009–17. 18

Muir AM, Hansen MJ, Bronte CR, Krueger CC (2016) If Arctic charr *Salvelinus alpinus* is 'the most diverse vertebrate', what is the lake charr *Salvelinus namaycush* ? *Fish and Fisheries*, **17**, 1194–1207. 2, 16

Muschick M, Indermaur A, Salzburger W (2012) Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. *Current Biology*, **22**, 2362–2368. 1

Nadeau NJ, Whibley A, Jones RT, *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 343–353. 17

Noakes DLG (2008) Charr truth: Sympatric differentiation in *Salvelinus* species. *Environmental Biology of Fishes*, **83**, 7–15. 2, 16

Norðdahl H, Ingólfsson O, Pétursson HG, Hallsdóttir M (2008) Late Weichselian and Holocene environmental history of Iceland. *Jökull*, **58**, 343–364. 15

Nosil P (2012) *Ecological Speciation*. Oxford series in ecology and evolution. Oxford University Press, Oxford ; New York. 1

Nugent CM, Easton AA, Norman JD, Ferguson MM, Danzmann RG (2017) A SNP Based Linkage Map of the Arctic Charr (*Salvelinus alpinus*) Genome Provides Insights into the Diploidization Process After Whole Genome Duplication. *G3; Genes|Genomes|Genetics*, **7**, 543–556. 4, 18

Nurse P, Thuriaux P (1980) Regulatory genes controlling mitosis in the fission yeast *Schizosaccharomyces pombe*. *Genetics*, **96**, 627–637. 18

Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M (2014) VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, **30**, 2076–2078. 7

Palsson A, Gibson G (2004) Association between nucleotide variation in *Egfr* and wing shape in *Drosophila melanogaster*. *Genetics*, **167**, 1187–1198. 18

Paradis E (2010) Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420. 7

Pardo-Diaz C, Salazar C, Baxter SW, *et al.* (2012) Adaptive Introgression across Species Boundaries in *Heliconius* Butterflies. *PLoS Genetics*, **8**, e1002752. 2

Parsons KJ, Sheets HD, Skúlason S, Ferguson MM (2011) Phenotypic plasticity, heterochrony and ontogenetic repatterning during juvenile development of divergent Arctic charr (*Salvelinus alpinus*). *Journal of Evolutionary Biology*, **24**, 1640–1652. 3

Parsons KJ, Skúlason S, Ferguson M (2010) Morphological variation over ontogeny and environments in resource polymorphic arctic charr (*Salvelinus alpinus*). *Evolution and Development*, **12**, 246–257. 3, 14

Pease JB, Haak DC, Hahn MW, Moyle LC (2016) Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biology*, **14**, e1002379. 2

Perreault-Payette A, Muir AM, Goetz F, *et al.* (2017) Investigating the extent of parallelism in morphological and genomic divergence among lake trout ecotypes in Lake Superior. *Molecular Ecology*, **26**, 1477–1497. 17

R Core Team (2015) R: A Language and Environment for Statistical Computing. 6, 7

Roberts A, Pachter L (2012) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, **10**, 71–73. 6

Saemundsson K (1992) Geology of the Thingvallavatn Area. *Oikos*, **64**, 40–68. 3, 15

Samy JKA, Mulugeta TD, Nome T, *et al.* (2017) SalmoBase: An integrated molecular data resource for Salmonid species. *BMC Genomics*, **18**, 482. 6

Sandlund OT, Gunnarsson K, Jónasson PM, *et al.* (1992) The arctic charr *Salvelinus alpinus* in Thingvallavatn. *Oikos*, **64**, 305–351. 4, 14, 16

Sandlund OT, Jonsson B, Malmquist HJ, *et al.* (1987) Habitat use of arctic charr *Salvelinus alpinus* in Thingvallavatn, Iceland. *Environmental Biology of Fishes*, **20**, 263–274. 3

Seehausen O, Butlin RK, Keller I, *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192. 2, 5, 17, 18

Seehausen O, Wagner CE (2014) Speciation in Freshwater Fishes. *Annual Review of Ecology, Evolution, and Systematics*, **45**, 621–651. 1, 3

Seppala M, Depew MJ, Martinelli DC, Fan CM, Sharpe PT, Cobourne MT (2007) *Gas1* is a modifier for holoprosencephaly and genetically interacts with sonic hedgehog. *Journal of Clinical Investigation*, **117**, 1575–1584. 18

Shapiro BJ, Leducq JB, Mallet J (2016) What Is Speciation? *PLOS Genetics*, **12**, e1005860. 17

Skúlason S, Antonsson T, Gudbergsson G, Malmquist HJ, Snorrason SS (1992) Varibility in Icelandic Arctic charr. *Icelandic Agricultural Sciences*, **6**, 143–153. 2

Skúlason S, Noakes DL, Snorrason SS (1989a) Ontogeny of trophic morphology in four sympatric morphs of arctic charr *Salvelinus alpinus* in Thingvallavatn, Iceland. *Biological Journal of the Linnean Society*, **38**, 281–301. 3, 14

Skúlason S, Smith TB (1995) Resource polymorphisms in vertebrates. *Trends in Ecology & Evolution*, **10**, 366–370. 2, 3

Skúlason S, Snorrason S, Noakes D, Ferguson M (1996) Genetic basis of life history variations among sympatric morphs of Arctic char *Salvelinus alpinus*. *Canadian Journal of Fisheries and Aquatic Sciences*, **53**, 1807–1813. 3

Skúlason S, Snorrason S, Ota D, Noakes D (1993) Genetically based differences in foraging behaviour among sympatric morphs of arctic charr (Pisces: Salmonidae). *Animal Behaviour*, **45**, 1179–1192. 3

Skúlason S, Snorrason SS, Noakes DLG, Ferguson MM, Malmquist HJ (1989b) Segregation in spawning and early life history among polymorphic Arctic charr, *Salvelinus alpinus*, in Thingvallavatn, Iceland. *Journal of Fish Biology*, **35**, 225–232. 3, 5, 16, 17

Smith TB, Skúlason S (1996) Evolutionary Significance of Resource Polymorphisms in Fishes, Amphibians, and Birds. *Annual Review of Ecology and Systematics*, **27**, 111–133. 2, 16

Snorrason SS, Skúlason S (2004) Adaptive Speciation in Northern Freshwater Fishes. In *Adaptive Speciation* (edited by U Dieckmann, M Doebeli, JAJ Metz, D Tautz), chap. 10, pp. 210–228. Cambridge University Press, Cambridge. 2, 3

Snorrason SS, Skúlason S, Sandlund OT, Malmquist HJ, Jonsson B, Jonasson PM (1989) Shape polymorphism in Arctic charr, *Salvelinus alpinus*. *Physiology and Ecology Japan*, **1**, 393–404. 11, 17, 19

Sutherland BJG, Gosselin T, Normandeau E, *et al.* (2016) Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. *Genome Biology and Evolution*, **8**, evw262. 4

Svavarsson E (2007) árangur í kynbótum á bleikju og næstu skref. In *Fræðaþing Landbúnaðarsins 4*, pp. 121–125. 14

Theis A, Ronco F, Indermaur A, Salzburger W, Egger B (2014) Adaptive divergence between lake and stream populations of an East African cichlid fish. *Molecular Ecology*, **23**, 5304–5322. 2

Timusk ER, Ferguson MM, Moghadam HK, Norman JD, Wilson CC, Danzmann RG (2011) Genome evolution in the fish family salmonidae: Generation of a brook charr genetic map and comparisons among charrs (Arctic charr and brook charr) with rainbow trout. *BMC genetics*, **12**, 68. 4

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**, 117–131. 3

Via S (2001) Sympatric speciation in animals: The ugly duckling grows up. *Trends in Ecology & Evolution*, **16**, 381–390. 1

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, **106**, 9939–9946. 1

Vijay N, Weissensteiner M, Burri R, Kawakami T, Ellegren H, Wolf JB (2017) Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. *Molecular Ecology*, **26**, 4284–4295. 2, 18

Vincent B, Dionne M, Kent MP, Lien S, Bernatchez L (2013) Landscape genomics in Atlantic salmon (*Salmo salar*): Searching for gene-environment interactions driving local adaptation. *Evolution*, **67**, 3469–3487. 17

Volpe JP, Ferguson MM (1996) Molecular genetic examination of the polymorphic Arctic charr *Salvelinus alpinus* of Thingvallavatn, Iceland. *Molecular ecology*, **5**, 763–72. 3, 11, 14, 15, 17, 19

Wilson AJ, Gíslason D, Skúlason S, *et al.* (2004) Population genetic structure of Arctic charr, *Salvelinus alpinus* from northwest Europe on large and small spatial scales. *Molecular ecology*, **13**, 1129–42. 2, 16

Wolf JBW, Ellegren H (2016) Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, **18**, 87–100. 2, 5

Woods PJ, Skúlason S, Snorrason SS, Kristjánsson BK, Malmquist HJ, Quinn TP (2012a) Intraspecific diversity in Arctic charr, *Salvelinus alpinus*, in Iceland: I. Detection using mixture models. *Evolutionary Ecology Research*, **14**, 973–992. 16

Woods PJ, Skúlason S, Snorrason SS, Kristjánsson BK, Malmquist HJ, Quinn TP (2012b) Intraspecific diversity in Arctic charr, *Salvelinus alpinus*, in Iceland: II. Which environmental factors influence resource polymorphism in lakes? *Evolutionary Ecology Research*, **14**, 993–1013. 2

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic Characterization and Linkage Disequilibrium Estimation of a Global Maize Collection Using SNP Markers. *PLoS ONE*, **4**, e8451. 12

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome biology*, **11**, R14. 7

Zimmerman E, Palsson A, Gibson G (2000) Quantitative trait loci affecting components of wing shape in *Drosophila melanogaster*. *Genetics*, **155**, 671–83. 18

# 7 Supplement

## 7.1 Supp-Tables and files

Table S1: Tab-delimited text file with the position, alleles, alternative frequencies within morphs, F-statistics between morphs and samples and predicted effect of the variant on protein composition for all the variants after our final filtering step.

In Dropbox: variants.tsv

**id:** Identifier for each variant from 1 to 19,252
**transcript_id:** Name of the transcript from the Trinity assembly.
**start:** Start position of polymorphism within contig.
**end:** End position of polymorphism within contig.
**ref:** Base(s) of the reference allele.
**alt:** Base(s) of the alternative allele.
**var_pos:** Position of polymorphism within contig.
**NCBI_id:** Sequence identifier for chromosome or scaffold in genome.
**Chr:** Name of chromosome.
**Chr_pos:** Position mapped to in chromosome.
**Freq_LB:** Frequency of alternative allele within the LB morph.
**Freq_PL:** Frequency of alternative allele within the PL morph.
**Freq_SB:** Frequency of alternative allele within the SB morph.
**Fst:** Estimate of $F_{ST}$ between morphs.
**Fpt:** Estimate of $F_{PT}$ or variation among samples (pools). See methods for further explanation.
**effect:** Which effect does the mutation have on the gene, 3'- and 5' UTR indicate mutations in those regions, synonymous do not change the reading frame but mutations that do so are coded as transitions from one amino acid to another (single letter a.a. code)
**gene_name:** Gene name based on SalmonBase annotation.

Table S2: VCF-file with the final set of variants after all filtering steps.

In Dropbox: var_freebayes_final.vcf

Table S3: Variants only found within one morph. Position within contig and genome, alleles, frequency within the morph, $F_{ST}$-values and predicted effect are shown.

| Transcript id | Var pos | Ref | Alt | Chr | Chr pos | Freq | $F_{ST}$ | $F_{PT}$ | Effect | Gene name |
|---|---|---|---|---|---|---|---|---|---|---|
| **Private to LB** | | | | | | | | | | |
| T143531c2g1i1 | 545 | G | A | LG2 | 1,668,110 | 0.33 | 0.19 | 0.32 | | CIGSSA_083607.t3 |
| T147645c2g1i1 | 571 | T | A | LG18 | 49,590,368 | 0.34 | 0.24 | 0.39 | | AN1-type zinc finger protein 4-like, isoform X1 |
| T149060c0g1i1 | 83 | T | A | LG4q.1:29 | 13,785,703 | 0.46 | 0.16 | 0.41 | UTR | Prenylcysteine oxidase |
| T149651c0g1i1 | 420 | C | G | LG18 | 47,847,696 | 0.34 | 0.21 | 0.43 | UTR | 40S ribosomal protein S24, isoform X3 |
| T161481c0g1i2 | 203 | C | T | LG13 | 50,967,669 | 0.45 | 0.30 | 0.68 | UTR | |
| T161641c2g2i1 | 1119 | G | A | LG23 | 40,674,682 | 0.37 | 0.24 | 0.34 | UTR | Ubiquitin-conjugating enzyme E2 D2\|predicted protein |
| T170659c2g1i1 | 434 | T | A | LG20 | 55,172,526 | 0.34 | 0.19 | 0.40 | | Low-density lipoprotein receptor-related protein 6 \|hypothetical weakly similar |
| T171446c2g1i3 | 159 | C | G | | | 0.48 | 0.15 | 0.55 | | |
| T174484c7g1i1 | 742 | C | T | LG23 | 12,588,360 | 0.32 | 0.23 | 0.40 | Syn | RNA-binding protein Nova-1-like, isoform X1 |
| T174834c7g3i2 | 629 | G | C | LG4p | 9,417,520 | 0.38 | 0.25 | 0.60 | T to R | TRPM8 channel-associated factor homolog, isoform X2 |
| T175088c0g3i1 | 698 | T | C | LG18 | 9,241,944 | 0.33 | 0.19 | 0.46 | UTR | ADP-ribosylation factor 1 like |
| T175804c1g1i1 | 766 | C | G | LG31 | 15,851,598 | 0.32 | 0.17 | 0.31 | R to G | protein FAM46A-like |
| T193474c0g1i1 | 1331 | G | A | LG14 | 38,002,543 | 0.33 | 0.19 | 0.46 | Syn | Cytochrome c-type heme lyase |
| **Private to PL** | | | | | | | | | | |
| T44119c0g1i1 | 152 | A | C | LG17 | 8,115,846 | 0.47 | 0.31 | 0.48 | | Glutamine–tRNA ligase |
| T70707c0g1i1 | 1085 | T | C | LG14 | 14,039,993 | 0.47 | 0.33 | 0.48 | UTR | Costars family protein ABRACL |
| T70707c0g1i1 | 1506 | T | C | LG14 | 14,040,414 | 0.43 | 0.31 | 0.45 | UTR | Costars family protein ABRACL |
| T83913c0g1i1 | 419 | G | A | LG18 | 59,124,877 | 0.55 | 0.38 | 0.67 | UTR | Adenosine kinase |
| T83913c0g1i1* | 1075 | C | T | LG18 | 59,122,320 | 0.66 | 0.60 | 0.73 | D to N | Adenosine kinase |
| T106859c0g1i1 | 366 | G | T | LG14 | 14,052,297 | 0.46 | 0.25 | 0.47 | | beta-taxilin, isoform X2 |
| T140173c0g1i2 | 148 | C | T | Scf2193 | 34,736 | 0.41 | 0.35 | 0.38 | UTR | N-myc proto-oncogene protein-like, isoform X2 |
| T148137c3g1i2 | 1158 | G | C | Scf16416 | 190,147 | 0.45 | 0.39 | 0.56 | UTR | dolichyldiphosphatase 1-like |
| T148137c3g1i2 | 1432;1440 | G(N7)T | T(N7)A | Scf16416 | 189,864 | 0.44 | 0.30 | 0.50 | UTR | dolichyldiphosphatase 1-like |
| T154282c0g1i1 | 142 | C | G | LG17 | 8,115,335 | 0.44 | 0.32 | 0.48 | | Glutamine–tRNA ligase |
| T154282c0g1i1 | 252;292;304 | G(N39)A -(N11)A | A(N39)T -(N11)C | LG17 | 8,115,195 | 0.52 | 0.37 | 0.58 | | Glutamine–tRNA ligase |
| T154282c0g1i1 | 575 | G | T | LG17 | 8,114,902 | 0.41 | 0.28 | 0.43 | | Glutamine–tRNA ligase |
| T156040c0g1i1 | 747 | C | A | LG4q.1:29 | 23,198,437 | 0.37 | 0.22 | 0.46 | UTR | Vacuolar protein sorting-associated protein 29 |
| T159843c0g1i1 | 1331 | G | A | LG31 | 24,512,078 | 0.40 | 0.26 | 0.39 | UTR | chromobox protein homolog 3-like, isoform X1 |
| T159843c0g1i1 | 1796 | G | T | LG31 | 24,512,546 | 0.58 | 0.44 | 0.55 | UTR | chromobox protein homolog 3-like, isoform X1 |
| T161458c2g1i1 | 77 | A | G | LG6.2 | 13,677,880 | 0.34 | 0.22 | 0.54 | | cyclin A2, isoform X2 |
| T161984c0g1i1 | 618 | C | T | LG31 | 22,464,046 | 0.33 | 0.20 | 0.33 | Syn | Trafficking protein particle complex subunit 3, isoform X1 |
| T162197c14g1i1 | 1353 | G | T | LG7 | 16,955,663 | 0.53 | 0.39 | 0.54 | Syn | Protein LZIC |
| T163256c7g7i3 | 3892 | T | A | Scf1703 | 56,973 | 0.54 | 0.40 | 0.51 | UTR | CIGSSA_123112.t11 |
| T163256c7g7i3 | 4065 | A | C | Scf1703 | 56,800 | 0.47 | 0.34 | 0.56 | UTR | CIGSSA_123112.t11 |
| T163989c5g1i1 | 1293 | T | C | LG7 | 27,364,962 | 0.36 | 0.11 | 0.52 | UTR | fibrous sheath CABYR-binding protein-like, isoform X1 |
| T165162c16g1i1 | 820 | G | A | LG1 | 17,578,711 | 0.34 | 0.15 | 0.26 | UTR | TOX high mobility group box family member 2-like, isoform X4 |
| T165162c16g1i1 | 974 | C | T | LG1 | 17,578,865 | 0.32 | 0.14 | 0.31 | UTR | TOX high mobility group box family member 2-like, isoform X4 |

Continued on Next Page…

Table S3: Variants only found within one morph. Position within contig and genome, alleles, frequency within the morph, $F_{ST}$-values and predicted effect are shown.

| Transcript id | Var pos | Ref | Alt | Chr | Chr pos | Freq | $F_{ST}$ | $F_{PT}$ | Effect | Gene name |
|---|---|---|---|---|---|---|---|---|---|---|
| T165162c16g1i1 | 975;984 | A(N8)A | T(N8)C | LG1 | 17,578,870 | 0.33 | 0.13 | 0.38 | UTR | TOX high mobility group box family member 2-like, isoform X4 |
| T168092c2g1i5 | 2724;2726 | TTA | ATT | LG7 | 26,670,845 | 0.35 | 0.23 | 0.51 | UTR | 45 kDa calcium-binding protein |
| T168398c0g1i4 | 1759 | G | T | LG32 | 20,522,655 | 0.33 | 0.28 | 0.35 | P to H | Kinesin-like protein |
| T168651c0g1i1 | 863 | A | G | LG1 | 17,539,258 | 0.36 | 0.30 | 0.40 | | alpha-1-syntrophin-like, isoform X1 |
| T168962c1g1i2 | 994 | G | C | LG14 | 7,560,825 | 0.35 | 0.24 | 0.38 | S to T | translocating chain-associated membrane protein 2-like |
| T169544c12g4i1 | 442 | A | C | LG41 | 294,246 | 0.48 | 0.34 | 0.53 | UTR | CIGSSA_056585.t2 |
| T170480c1g1i1 | 2202 | A | T | LG4q.2 | 9,848,462 | 0.34 | 0.16 | 0.49 | | CIGSSA_044851.t3 |
| T170529c17g1i1 | 176 | A | T | LG36 | 28,810,075 | 0.40 | 0.31 | 0.49 | | integral membrane protein 2B-like, isoform X1 |
| T171302c3g1i1 | 521 | C | T | LG17 | 8,116,418 | 0.42 | 0.24 | 0.39 | | Glutamine–tRNA ligase |
| T171326c3g1i2 | 1206 | A | T | LG11 | 21,367,985 | 0.33 | 0.15 | 0.48 | UTR | Peptidylarginine deiminase hypothetical |
| T171686c7g2i1 | 193 | G | C | LG17 | 8,312,053 | 0.49 | 0.39 | 0.50 | | Zgc:110709 |
| T171686c7g2i1 | 665 | G | C | LG17 | 8,311,589 | 0.53 | 0.42 | 0.58 | | Zgc:110709 |
| T171686c7g2i1 | 982 | T | A | LG17 | 8,311,272 | 0.50 | 0.43 | 0.50 | | Zgc:110709 |
| T171686c7g2i1 | 1424 | T | G | LG17 | 8,310,831 | 0.42 | 0.33 | 0.46 | | Zgc:110709 |
| T172162c0g1i1 | 425 | T | A | LG7 | 16,736,407 | 0.47 | 0.35 | 0.54 | | TAR DNA binding protein |
| T172217c4g4i1 | 1238 | T | A | LG32 | 6,053,989 | 0.53 | 0.36 | 0.57 | UTR | WW domain-containing adapter protein with coiled-coil |
| T172217c4g4i1 | 2113 | A | G | LG32 | 6,054,862 | 0.58 | 0.47 | 0.58 | UTR | WW domain-containing adapter protein with coiled-coil |
| T172357c4g1i1 | 2197 | A | C | | | 0.42 | 0.41 | 0.38 | Syn,UTR | coronin-7-like |
| T172967c3g1i1 | 199 | G | C | LG35 | 8,330,659 | 0.51 | 0.35 | 0.51 | | RAB11a member RAS oncogene family |
| T172992c11g4i1 | 448 | T | C | LG14 | 14,078,902 | 0.57 | 0.44 | 0.60 | Syn | Beta-taxilin\|hypothetical |
| T172933c13g1i1 | 2089 | C | T | LG7 | 16,959,001 | 0.48 | 0.33 | 0.47 | | calsyntenin 1, isoform X4 |
| T173769c10g1i2 | 742;743;748 | TATTTTT | ATTTTTG | Scf1321 | 226,642 | 0.34 | 0.18 | 0.49 | | Inosine-5-monophosphate dehydrogenase 1 |
| T174499c6g3i3 | 1150 | A | G | LG7 | 16,917,732 | 0.50 | 0.38 | 0.53 | S to G | ubiquitination factor E4B, UFD2 homolog (S. cerevisiae), isoform X1 |
| T174814c3g1i1 | 411 | T | A | LG4q.1:29 | 37,790,310 | 0.33 | 0.24 | 0.47 | | CMP-N-acetylneuraminate-poly-alpha-28-sialyltransferase |
| T174821c6g1i1 | 967 | C | G | LG36 | 31,081,180 | 0.41 | 0.35 | 0.40 | | unconventional myosin-Ib-like, isoform X1 |
| T174821c6g1i1 | 1199 | G | A | LG36 | 31,080,948 | 0.37 | 0.23 | 0.36 | | unconventional myosin-Ib-like, isoform X1 |
| T175009c8g1i2 | 2905 | C | T | LG30 | 8,090,556 | 0.47 | 0.34 | 0.45 | UTR | ATP-binding cassette sub-family F member 1-like |
| T175236c1g3i4 | 5881 | T | G | LG31 | 24,323,360 | 0.61 | 0.45 | 0.56 | UTR | collagen alpha-1(XVI) chain-like, isoform X1 |
| T175348c1g10i1 | 258 | A | G | LG20 | 21,264,718 | 0.32 | 0.18 | 0.40 | | ubiquitin domain containing 2 |
| T175403c90g7i1 | 1245 | A | C | Scf396 | 147,836 | 0.73 | 0.55 | 0.83 | UTR | peptidyl-prolyl cis-trans isomerase FKBP3-like |
| T175479c1g2i5 | 407 | A | C | LG35 | 8,345,621 | 0.50 | 0.38 | 0.50 | V to G | Ubiquilin-4 |
| T176050c1g1i1 | 5019 | C | A | LG35 | 8,316,952 | 0.52 | 0.44 | 0.51 | UTR | RNA-binding protein MEX3A |
| T210338c0g1i1 | 1772 | G | A | LG17 | 8,125,541 | 0.50 | 0.36 | 0.48 | Syn | glutaminyl-tRNA synthetase |
| **Private to SB** | | | | | | | | | | |
| T145511c0g1i1 | 2002 | T | C | Scf4114 | 11,937 | 0.26 | 0.26 | 0.25 | Syn | Myelin expression factor 2\|hypothetical |
| T160725c0g1i2 | 919 | A | G | LG17 | 22,513,650 | 0.40 | 0.41 | 0.37 | Syn | transcription factor 15-like |
| T162598c0g1i1 | 729;742 | G(N12)C | A(N12)T | Scf1046 | 1,457 | 0.44 | 0.23 | 0.44 | | chromobox homolog 6 |
| T165516c2g3i4 | 382 | C | A | Scf4168 | 5,044 | 0.27 | 0.19 | 0.46 | UTR | |
| T165516c2g3i4 | 873;876;912; 924;927 | CCAC(N35)A -(N10)AAAT | TCAT(N35)T -(N10)GAAC | LG18 | 24,320,993 | 0.43 | 0.38 | 0.65 | Syn | |
| T168835c3g2i2 | 1142 | T | A | LG4q.1:29 | 21,965,763 | 0.33 | 0.33 | 0.37 | UTR | Phosphoglycerate mutase 2 |
| T169028c43g1i1 | 3476 | A | C | LG32 | 2,199,330 | 0.31 | 0.29 | 0.40 | UTR | Muscle-related coiled-coil protein |

Continued on Next Page…

Table S3: Variants only found within one morph. Position within contig and genome, alleles, frequency within the morph, $F_{ST}$-values and predicted effect are shown.

| Transcript id | Var pos | Ref | Alt | Chr | Chr pos | Freq | $F_{ST}$ | $F_{PT}$ | Effect | Gene name |
|---|---|---|---|---|---|---|---|---|---|---|
| T169667c0g1i4 | 2617 | A | G | LG18 | 12,429,904 | 0.28 | 0.30 | 0.32 | F to S | DNA topoisomerase 2-alpha |
| T169667c0g1i4 | 2835 | G | A | LG18 | 12,430,772 | 0.28 | 0.31 | 0.32 | Syn | DNA topoisomerase 2-alpha |
| T169667c0g1i4 | 3054 | G | T | LG18 | 12,431,454 | 0.28 | 0.30 | 0.30 | Syn | DNA topoisomerase 2-alpha |
| T169667c0g1i4 | 3198 | A | G | LG18 | 12,431,726 | 0.26 | 0.30 | 0.30 | Syn | DNA topoisomerase 2-alpha |
| T169667c0g1i4 | 3375 | A | G | LG18 | 12,432,251 | 0.29 | 0.31 | 0.31 | Syn | DNA topoisomerase 2-alpha |
| T169667c0g1i4 | 3447 | G | A | LG18 | 12,432,602 | 0.29 | 0.31 | 0.36 | Syn | DNA topoisomerase 2-alpha |
| T169667c0g1i4 | 3672 | T | C | LG18 | 12,433,037 | 0.28 | 0.32 | 0.32 | Syn | DNA topoisomerase 2-alpha |
| T169667c0g1i4 | 4308 | C | A | LG18 | 12,435,579 | 0.27 | 0.29 | 0.32 | Syn | DNA topoisomerase 2-alpha |
| T170294c11g5i4 | 1744 | T | C | LG17 | 22,247,612 | 0.33 | 0.31 | 0.32 | Syn | ras-related protein Rap-1A-like, isoform X1 |
| T173826c6g1i1 | 342 | A | G | LG28 | 18,758,190 | 0.36 | 0.32 | 0.43 | | Si:dkey-77p13.2 |
| T210331c21g1i1 | 4709 | G | A | MG | 3,411 | 0.74 | 0.68 | 0.79 | UTR | |

*Validated in KASP-assay

Table S4: Gene ontology categories enriched in variants private to PL-charr ($p < 0.01$ in the other morphs) on the transcripts ($tr$) and gene ($ge$) level. The number of transcripts and genes with observed private variants ($PL_{tr}$ and $PL_{ge}$) and the total number of transcripts and genes tested in the category ($Tot_{tr}$ and $Tot_{ge}$) are shown. The multiple testing corrected p-value or false discovery rate ($FDR$) is also shown for both levels.

| Category | Term | $PL_{tr}$ | $Tot_{tr}$ | $FDR_{tr}$ | $PL_{ge}$ | $Tot_{ge}$ | $FDR_{ge}$ |
|---|---|---|---|---|---|---|---|
| GO:0006425 | glutaminyl-tRNA aminoacylation | 4 | 4 | 2.47e-06 | 1 | 1 | 1.0000 |
| GO:0006433 | prolyl-tRNA aminoacylation | 4 | 4 | 2.47e-06 | 1 | 1 | 1.0000 |
| GO:0006424 | glutamyl-tRNA aminoacylation | 4 | 6 | 2.44e-05 | 1 | 3 | 1.0000 |

Table S5: Genetic polymorphisms studied in the population sample of Arctic charr.

In Dropbox: Supplement_KaspMarkersSequencesv2.ods

**Transcript_ID:** Name of the transcript from the Trinity assembly
**Variant:** Name of variant.
**Gene:** Short name of gene.
**Gene_name:** Long name of gene.
**Start:** Position of polymorphism within contig.
**Ref:** Base(s) of the reference allele.
**Alt:** Base(s) of the alternative allele.
**Outg:** Base(s) in the salmonid outgroup.
**Chr_NCBI_id:** Sequence identifier for chromosome or scaffold in genome.
**Chr:** Name of chromosome.
**Chr_position:** Position mapped to in chromosome.
**Effect:** Which effect does the mutation have on the gene, 3'- and 5' UTR indicate mutations in those regions, synonymous do not change the reading frame but mutations that do so are coded as transitions from one amino acid to another (single letter a.a. code)
**Marker_ID:** Abbreviated ID of each marker.
**Sequence:** The sequence used to design the KASP assay, the polymorphism are marked by e.g. [A/T].

Table S6: Datafile of the genotypes for the 22 markers scored in the population samples from the four sympatric charr morphs (coded by bases, missing data indicated by "NA"),

In Dropbox: Datafile_ResultSameinad7.csv

Table S7: Estimates of F-statistics for the entire KASP data and tests of Hardy-Weinberg proportions for the entire KASP dataset (Total) and individual morphs.

In Dropbox: Supplement_KaspFstHW.ods

**Gene_SNP:** Name of variant
**Total (P):** Significance of the test for Hardy Weinberg proportions as estimated from Fisher's exact tests, on the entire dataset
**LB (P):** Significance of the test for Hardy Weinberg proportions as estimated from Fisher's exact tests for LB
**PI (P):** Significance of the test for Hardy Weinberg proportions as estimated from Fisher's exact tests for PI
**PL (P):** Significance of the test for Hardy Weinberg proportions as estimated from Fisher's exact tests for PL
**SB (P):** Significance of the test for Hardy Weinberg proportions as estimated from Fisher's exact tests for SB
**Fst:** F-statistics for variation between morphs
**Fit:** F-statistics for variation between individuals
**Fis:** F-statistics for variation within population

Table S8: Datafile of the LD ($r^2$) for all pairs of markers, by morph

In Dropbox: SuppLDAll.ods

## 7.2 Supp-Figures

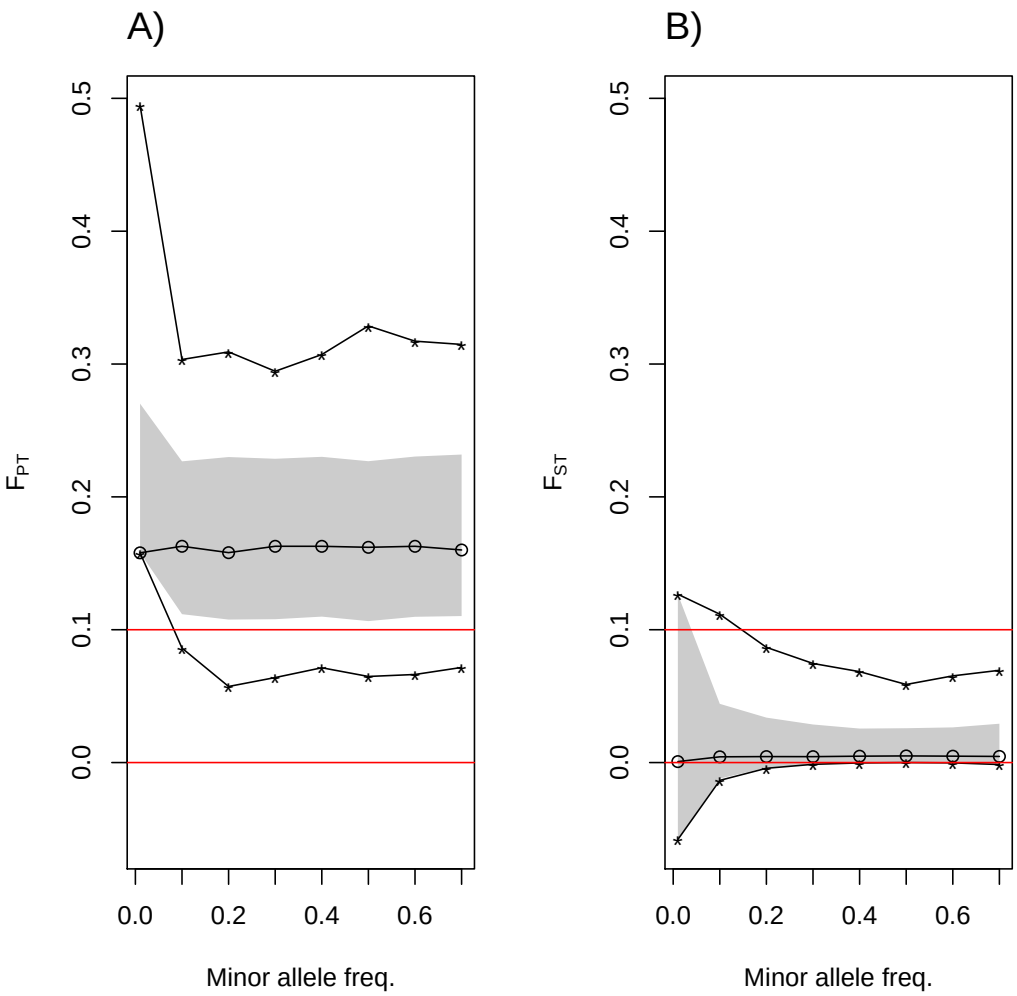


Figure S1: Results from simulations on F-statistics for a biallelic variant with the same allele frequency among morphs. The gray area indicate the 95% confidence area. The median, minimum and maximum values are shown by dots and lines. A) Shows $F_{PT}$ values and B) $F_{ST}$ values. As expected $F_{PT}$ values are high and we chose 0.1 (red vertical line) as cutoff as it is outside the 95% confidence area in the simulations.

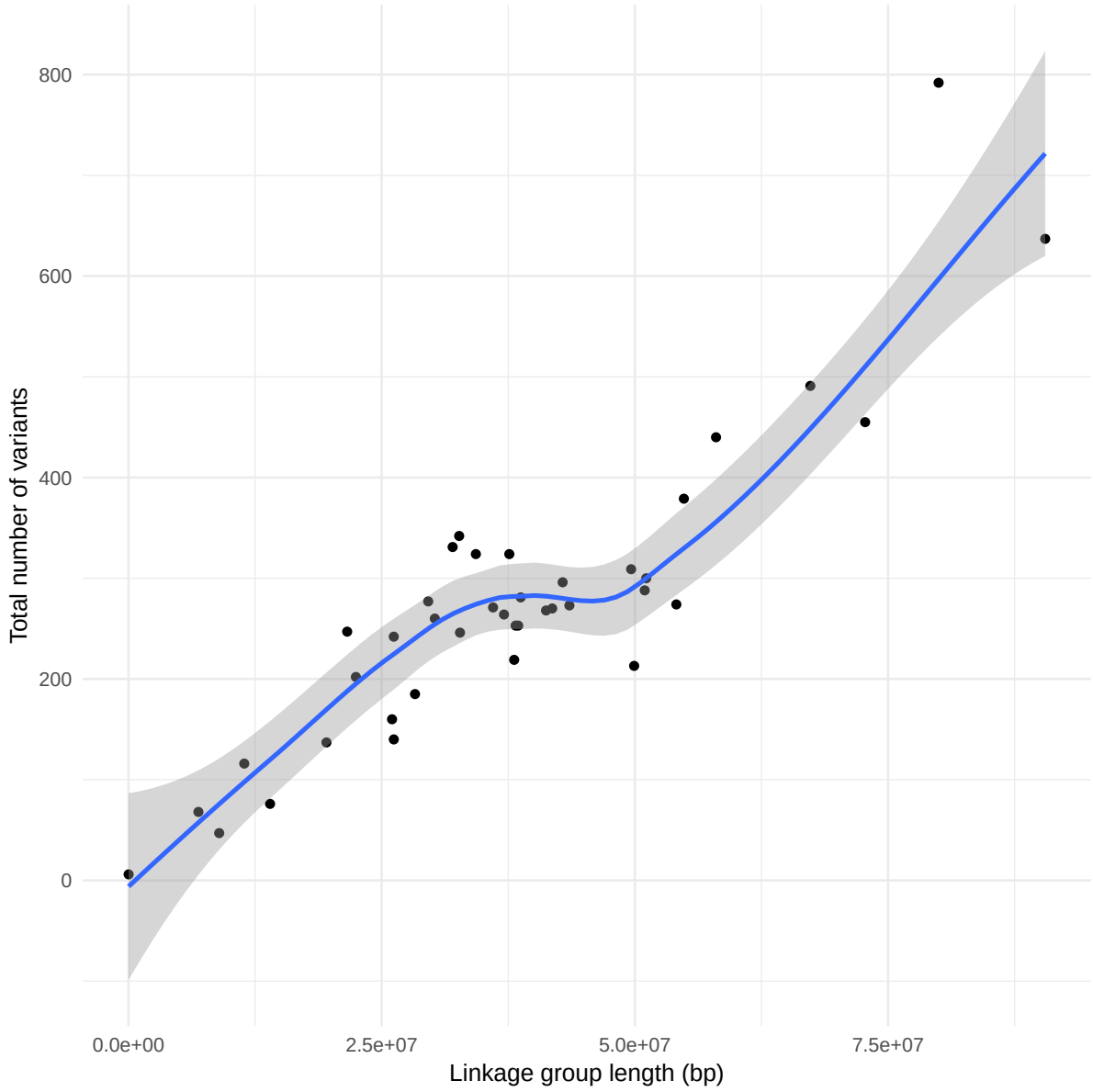


Figure S2: Total number of transcriptome variants by linkage group length. A loess smooth curve (blue) with 95% confident interval (gray) is also graphed.

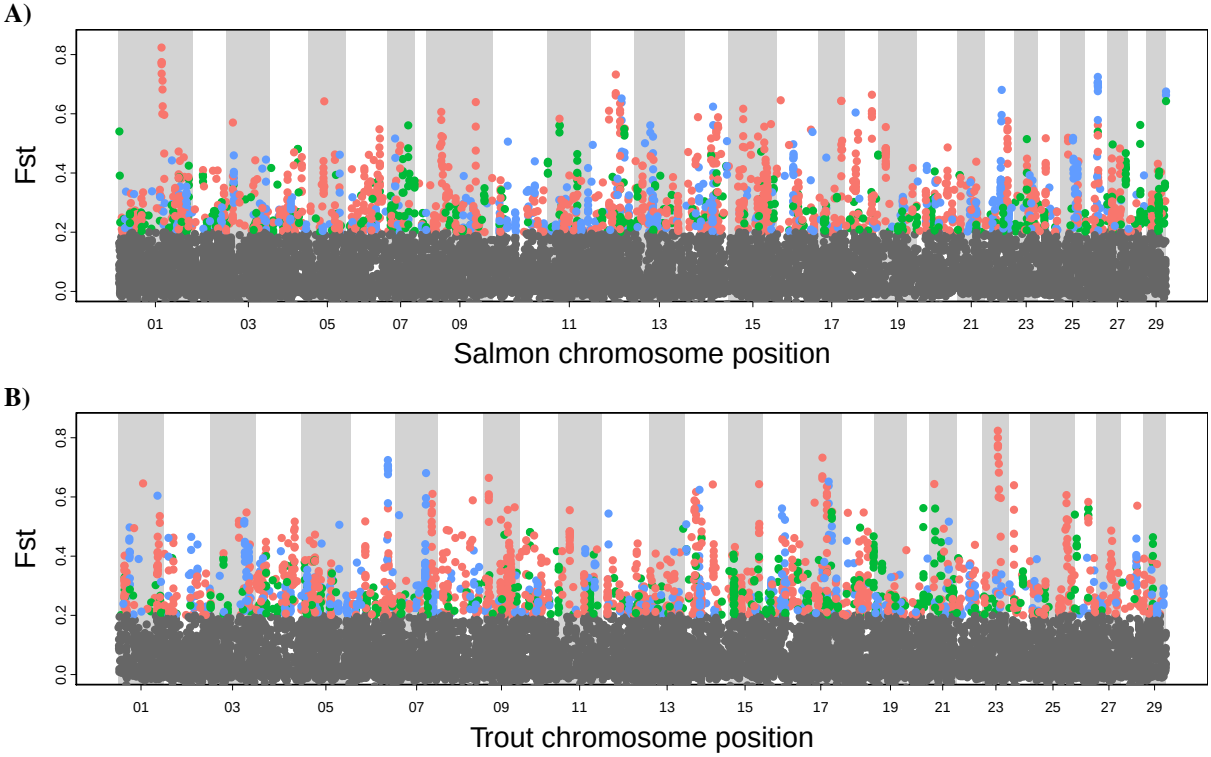


Figure S3: $F_{ST}$ values plotted by position of variants on the salmon **A)** and rainbow trout **B)** genome. The colors indicate which morphs differs most strongly in allele frequency from the other two for variants with $F_{ST}$ >0.2. Red PL, blue SB and green LB

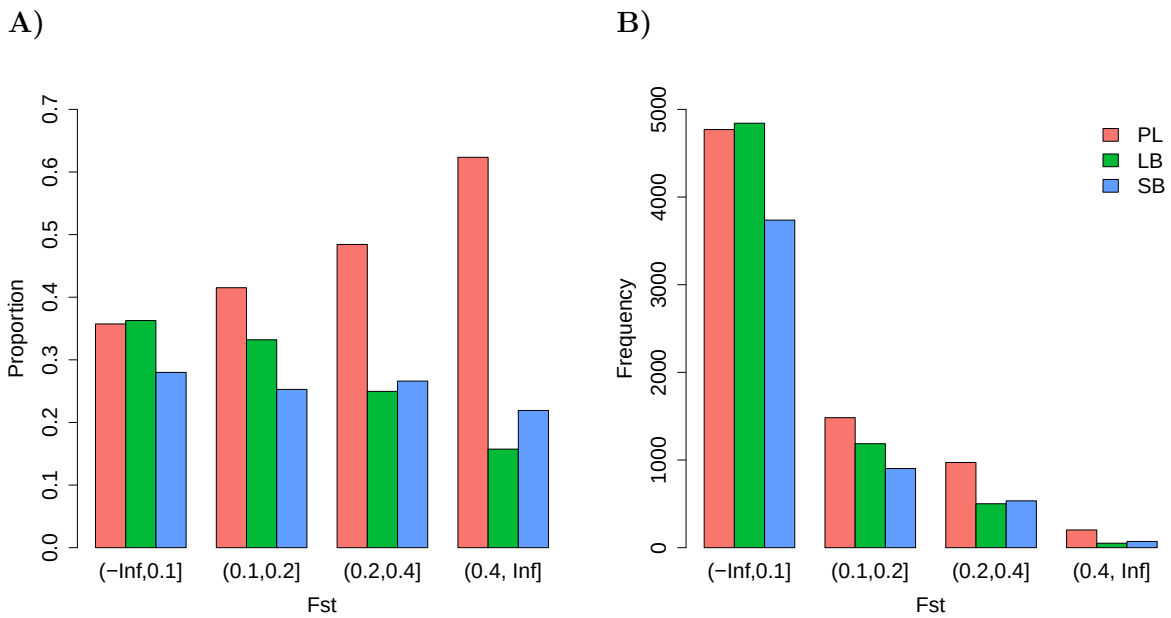


Figure S4: **A)** Proportion of variants (within $F_{ST}$ groups) for different $F_{ST}$-values grouped by the morph with highest deviation in allele frequency from the other two. **B)** The number of variants in each $F_{ST}$ category grouped by the morph with highest deviation in allele frequency from the other two. The legend in **B)** also applies to **A)**

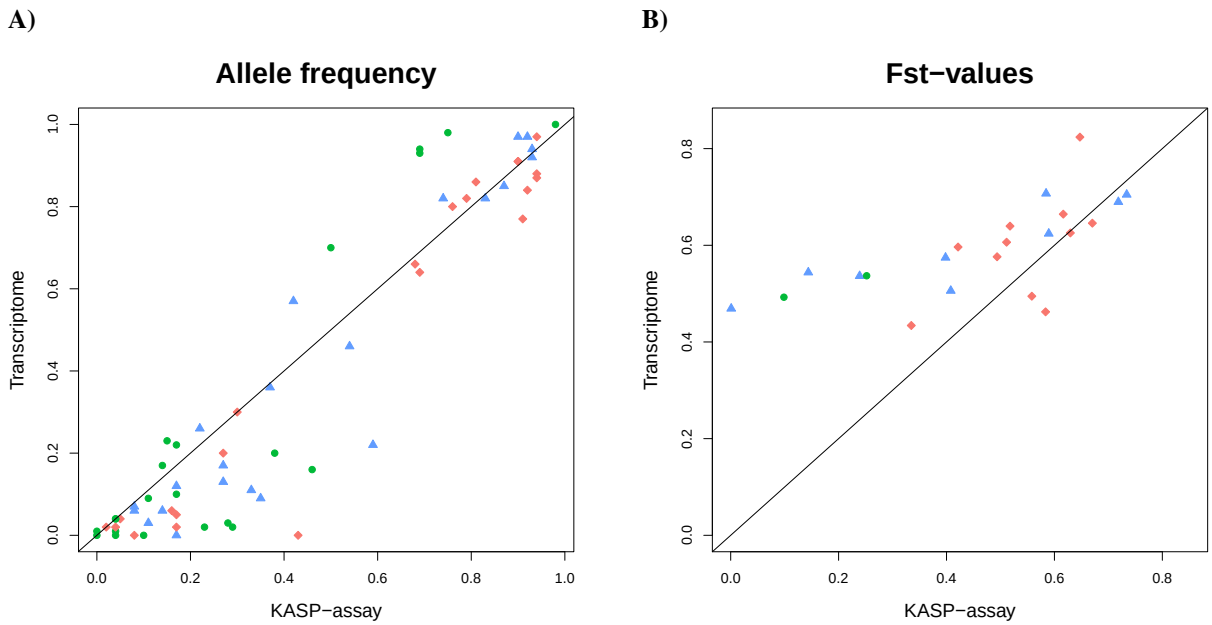


Figure S5: Tight relationship was found between the allele frequencies and $F_{ST}$ values estimated from the transcriptome and the population genetic sample (Kasp assay). **A)** The allele frequencies for the 22 markers were estimated for each morph (color coded) and the estimates from the two methods show high positive correlation (Kendall's $\tau = 0.71$, Pearson's $r = 0.94$, $p < 0.0001$). The diagonal line represents the 1:1 relationship. **B)** $F_{ST}$-values calculated for the transcriptome and the population sample had comparatively weaker association (Kendall's $\tau = 0.60$, Pearsons's $r = 0.67$, $p < 0.001$). Notably, five markers deviated from the 100% relationship (diagonal line), with higher $F_{ST}$ in the transcriptome compared to the population sample (due to underestimation of some rarer allele frequencies in the transcriptome). The colors indicate which morphs shows the highest deviation from the other two in mean allele frequency for each marker in the transcriptome.

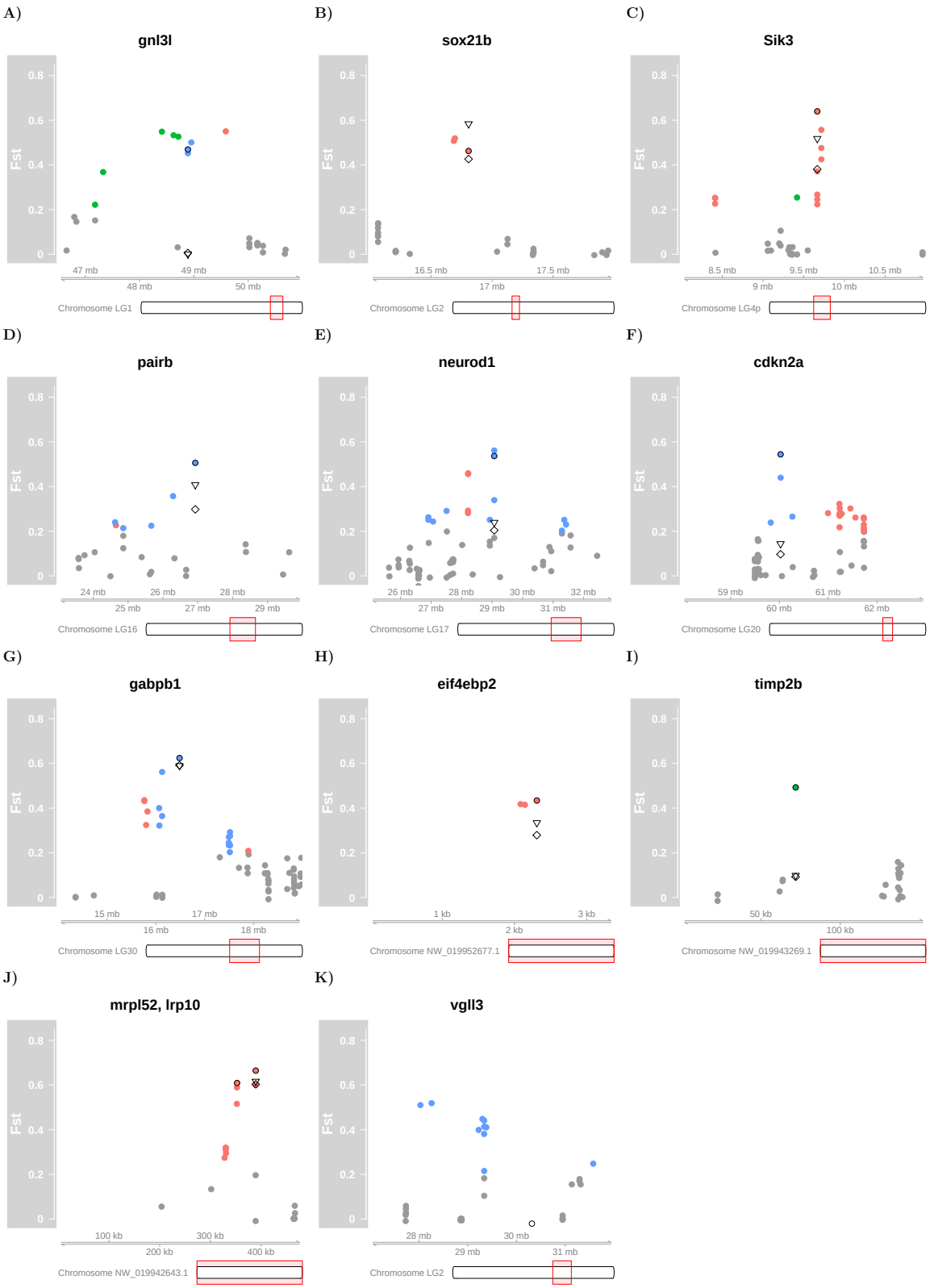


Figure S6: **A-J)** Detailed view of chromosomal regions of variants in KASP-assay and nearby transcriptome variants (not-validated). The colored dots mark $F_{ST}$ values from the transcriptome as in Fig. 4 and the variants taken for validation are marked with a black circl (○). The triangles (▽) show the $F_{ST}$ value from the KASP-assay for the three transcriptome morphs (PL, SB and LB) and the diamonds (◇) $F_{ST}$ for all morphs (including PI). **K)** Detailed view of variants nearby the *vgll3* locus. The circle (○) represents the location of the *vgll3* gene, not transcribed and thus no variants detected, using the same color code for variants.