F1000Research

RESEARCH ARTICLE

The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs [version 1; referees: 2 approved with reservations]

Johannes Gudbrandsson[1], Ehsan P. Ahi[1], Sigridur R. Franzdottir[1],
Kalina H. Kapralova[1], Bjarni K. Kristjansson[2], S. Sophie Steinhaeuser[1],
Valerie H. Maier[1], Isak M. Johannesson[1], Sigurdur S. Snorrason[1],
Zophonias O. Jonsson[1], Arnar Palsson[1]

[1]Institute of Life and Environmental Sciences, University of Iceland, Reykjavik, 101, Iceland
[2]Holar University College, Saudarkrokur, 551, Iceland

Abstract

Species and populations with parallel evolution of specific traits can help illuminate how predictable adaptations and divergence are at the molecular and developmental level. Following the last glacial period, dwarfism and specialized bottom feeding morphology evolved rapidly in several landlocked Arctic charr *Salvelinus alpinus* populations in Iceland.
To study the genetic divergence between small benthic morphs and limnetic morphs, we conducted RNA-sequencing charr embryos at four stages in early development. We studied two stocks with contrasting morphologies: the small benthic (SB) charr from Lake Thingvallavatn and Holar aquaculture (AC) charr. The data reveal significant differences in expression of several biological pathways during charr development. There was also an expression difference between SB- and AC-charr in genes involved in energy metabolism and blood coagulation genes. We confirmed differing expression of five genes in whole embryos with qPCR, including *lysozyme* and *natterin-like* which was previously identified as a fish-toxin of a lectin family that may be a putative immunopeptide. We also verified differential expression of 7 genes in the developing head that associated consistently with benthic v.s.limnetic morphology (studied in 4 morphs). Comparison of single nucleotide polymorphism (SNP) frequencies reveals extensive genetic differentiation between the SB and AC-charr (~1300 with more than 50% frequency difference). Curiously, three derived alleles in the otherwise conserved 12s and 16s mitochondrial ribosomal RNA genes are found in benthic charr.

The data implicate multiple genes and molecular pathways in divergence of small benthic charr and/or the response of aquaculture charr to domestication. Functional, genetic and population genetic studies on more freshwater and anadromous populations are needed to confirm the specific loci and mutations relating to specific ecological traits in Arctic charr.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees	
	1	**2**
REVISED version 2 published 25 Apr 2016		
version 1 published 01 Jun 2015	? report	? report

1 **Daniel Macqueen**, University of Aberdeen UK

2 **Anne Dalziel**, Laval University Canada

Discuss this article

Comments (0)

Corresponding authors: Johannes Gudbrandsson (jog7@hi.is), Arnar Palsson (apalsson@hi.is)

Competing interests: No competing interests were disclosed.

Introduction

Historical contingencies and chance shape organisms during evolution[1,2], but convergence in phenotype and molecular systems indicates that evolution is to some extent predictable[3,4]. Identification of genes and variants that influence evolved differences is not a trivial task[5]. Ideal systems to study the role of chance and necessity in ecological evolution would be related species or populations with readily observable phenotypic variation, living in a tractable ecological setting, and most crucially showing parallel evolution of specific traits within/among species/populations. Examples of such a species complex are the finches of the Galapagos islands, cichlids in the African great lakes are exciting multi-species systems in this respect[6,7]. The threespine stickleback has also emerged as a model "single species" system[8]. The amount of diversity in the feeding specializations of fish provide great opportunities for studying adaptation and divergence at the developmental and genetic level.

Transcriptomic methods have been to address evolutionary and ecological questions in fish. For example microarrays were used to compare gene expression in anadromous and resident populations of brown trout (*Salmo trutta*), revealing that life history was a better predictor of gene expression in the liver than relatedness[9]. The newer technique, RNA-sequencing (RNA-seq) has been applied to species such as the Mexican cavefish (*Astyanax mexicanus*), cod (*Gadus morhua*) brook charr (*Salvelinus fontinalis*) and Atlantic salmon (*Salmo salar*)[10–15], addressing questions concerning evolution, molecular genetics, development and aquaculture. RNA-seq was used to study salinity tolerance in Arctic charr, linking expression and quantitative trait loci[16]. Microarray studies of adult lake whitefish (*Coregonus clupeaformis*) pointed to parallel expression differences between benthic and limnetic forms[17]. Filteau *et al.* (2013)[18] found that a set of coexpressed genes differentiated the two whitefish morphotypes, implicating Bone morphogenesis protein (BMP) signaling in the development of ecological differences in tropic morphology. One approach to identify pathways related to function or morphological differences is to study gene expression during development[19,20].

Some northern freshwater fish species exhibit frequent parallelism in trophic structures and life history and in several cases are they found as distinct resource morphs[8,21–25]. One of these species, Arctic charr (*Salvelinus alpinus*), is well suited for studying the developmental underpinnings of trophic divergence and parallel evolution. Local adaptation has been extensively studied in the salmonid family, to which Arctic charr belongs[26]. The family is estimated to be between 63.2 and 58.1 million years old[27,28]. A whole genome duplication event occurred before the radiation of the salmonid family[29–32] which has provided time for divergence of ohnologous genes (paralogous genes originated by whole genome duplication event). Furthermore, recent estimates from the rainbow trout (*Oncorhynchus mykiss*) genome suggest that ohnologous genes were lost at a rate of about 170 ohnologous genes per million years and by utilizing multiple data sources the genome assembly problem of this family can be solved[32]. *De novo* assembly of genomes and transcriptomes is complicated if many paralogs are present, such as in salmonids. Furthermore, for data with short reads, mapping to a related reference genome/transcriptome is recommended over *de novo* assembly[33].

Molecular studies of the highly polymorphic Arctic charr

Following the end of the last glacial period, about 10.000 years ago, Arctic charr colonized northern freshwater systems[34]. It can be found as anadromous or lake/stream residents and exhibits high level of within species polymorphism[23,34]. Resource polymorphism in charr correlates with ecological attributes[35–37]. For instance small charr with benthic morphology, are found in multiple lavaspring and pond habitats in Iceland[38], and a comparative study of Icelandic lakes[37] found that lakes with greater limnetic habitat, fewer nutrients, and greater potential for zooplankton consumption appeared to promote resource polymorphism. Some of the larger lakes contain two or more distinct morphs, typically limnetic and benthic forms. Multiple lines of evidence show that these differences stem both from environmental and genetic causes[39–43]. The best studied example of sympatric charr are the four morphs in Lake Thingvallavatn[44]; two have a benthic morphotype, a large benthivorous (LB-charr) and a small benthivorous (SB-charr), and two morphs are limnetic, a large piscivorous morph (PI-charr) and small planktivorous morph (PL-charr)[45]. Both PL and PI-charr operate in open water and feed on free-swimming prey, PL on planktonic crustaceans and PI on small fish.

Several population genetics studies, using allozymes and mtDNA revealed no differences among charr populations[46–48], while studies on microsatellite markers and nuclear genes, reveled both subtle[49–51] and strong genetic differences among morphs[52]. Importantly Kapralova *et al.* (2011)[51] concluded that small benthic morphs have evolved repeatedly in Iceland and that gene flow has been reduced between the PL and SB morphs in Lake Thingvallavatn since its formation approximately 10,000 years ago[53]. We also discovered genetic separation in immunological genes (*MHCIIα* and *cath2*) between morphs in Iceland and within the lake[52], consistent with ecologically driven evolution of immune functions. Recently qPCR analyses showed that expression of mTOR pathway components in skeletal muscle correlates with the SB-charr form in Iceland[54], but it is unknown whether there is genetic differentiation in those genes or upstream regulators.

Because individual genes have distinct histories[55,56], genome wide methods are needed to identify genes and mutation that associate with divergence. Icelandic aquaculture charr (AC) was founded with fish from the north of Iceland, and has been bred at Holar University College since 1990[57]. The Holar AC-charr has responded to artificial selection in growth and performance characteristics, and is now the dominant charr breed in aquaculture in Iceland. While clearly a derived form, it has retained general limnetic craniofacial morphotype (Figure 1). In this study we compare SB-charr from Lake Thingvallavatn and AC-charr because i) SB charr represents an extensively studied and derived form of charr, that has been separated from Anadromous fish for approx. 10,000 years, ii) of the availability of abundant AC material and iii) we wanted an extreme contrast, because of budget reasons we could only sequence 8 samples at the time. This rather extreme contrast is justified as the data and other studies[58,59] building on this data illustrate (see discussion).

The aims of this project are threefold. First, to find genes and pathways related to the development of phenotypic differences between benthic and limnetic Arctic charr morphs. Second, to screen for

Figure 1. The two contrasting morphs used in this study. Adult individuals of the two morphs; the Holar aquaculture charr above and the small benthic charr from Lake Thingvallavatn below. Differences in size, coloration and head morphology are apparent.

signals of genetic differentiation that may relate to divergence of benthic and limnetic charr. Third, we set out to verify a subset of the expression and genetic signals, in benthic and limnetic morphs. We conducted RNA-sequencing of developing offspring of two contrasting Arctic charr morphs, a small benthic charr from Lake Thingvallavatn and Icelandic aquaculture charr conforming to a limnetic morphotype. This identified candidate genetic changes and differential expression of developmental genes that may affect jaw and craniofacial traits which separate benthic and limnetic morphotypes in charr.

Methods
Sampling, rearing and developmental series
We set up crosses and reared embryos in the laboratory as previously described in 58. Embryos from four charr morphs were studied: an aquaculture charr (AC-charr) from the Holar breeding program[57] and three natural morphs from Lake Thingvallavatn; small benthivorous (SB), large benthivorous (LB) and small planktivorous (PL) charr[60]. Samples of the first two, AC and SB-charr, which exhibit contrasting adult size and morphology (Figure 1), were collected in 2009 and material sent for developmental transcriptome sequencing. The latter two were sampled in 2010 and used for qPCR and SNP studies of selected genes. Briefly, in September 2009 we got material from spawning AC-charr from the Holar breeding program[57] and spawning SB-charr collected via gill netting in Olafsdrattur in Lake Thingvallavatn. Similarly, in the 2010 spawning season SB-, LB- and PL-charr were collected from Lake Thingvallavatn. Fishing permissions were obtained from the Thingvellir National Park Commission and the owner of the Mjóanes farm. For each parent group, eggs from several females were pooled and fertilized using milt from several males from the same group. Embryos were reared at ~5°C under constant water flow and in complete darkness at the Holar University College experimental facilities in Verid, Saudárkrókur. The water temperature was recorded twice daily and the average was used to estimate the relative age of the embryos using tausomite units (τs)[61]. Embryos and juveniles were sampled at designated time points, placed in RNAlater (Ambion) and frozen at -20°C. Post hatching juveniles were reared at the same temperature on standard Aquaculture food. For the investigation of different tissues of adult aquaculture charr (AC) from Hólar (fish size 20–25 cm) were used. Six randomly selected individuals were killed (by cutting through spinal cord) and dissected, and samples were taken from the skin, heart, liver, gills, spleen, intestine and kidney of each fish. The samples were placed in RNAlater (Ambion) and stored at -20°C. DNA for population genetic analyses was taken from our previous study[52].

Fishing in Lake Thingvallavatn was with permissions obtained both from the owner of the land in Mjóanes and from the Thingvellir 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 157/2012). Sampling was performed by Holar University College Aquaculture Research Station (HUC-ARC) personnel. HUC-ARC has an operational license according to Icelandic law on aquaculture (Law 71/2008), which includes clauses of best practices for animal care and experiments.

RNA extraction and transcriptome sequencing
Embryos of AC- and SB-charr sampled in 2009 were used for transcriptome sequencing. For this we focused on the time covering development of pharyngeal arches and morphogenesis of the head: at 141, 163, 200 and 433 τs (post fertilization). For each combination of morphs and timepoints we pooled RNA from approximately six individuals. RNA extraction and following steps were performed as described earlier[58,62]. The embryos were dechorionated and homogenized with a disposable Pellet Pestle Cordless Motor tissue grinder (Kimble Kontes, Vineland, NJ, USA) and RNA was extracted into two size-fractions using the Ambion mirVana kit (Life Technologies, Carlsbad, CA, USA). The high molecular weight fraction was further used for mRNA-seq and RNA quality was analysed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). First and second strand cDNA synthesis, fragmentation, adapter ligation and amplification were performed using the mRNA-Seq 8-Sample Prep Kit (Illumina, San Diego, CA, USA) according to manufacturer's instructions. Sequencing was

performed at DeCode genetics (Reykjavík, Iceland) using SOL-EXA GAII technology (Illumina, San Diego, CA, USA). The sequencing reads were deposited into the NCBI SRA archive under BioProject identifier PRJNA239766 and with accession numbers: SRX761559, SRX761571, SRX761575, SRX761577, SRX761451, SRX761461, SRX761490 and SRX761501.

The embryos sampled in 2010 were used for qPCR expression analyses. RNA was extracted from six whole embryos, in two replicates (two repetitions X three fish) (AC and SB sampled at 161 and 200 τs). For the extraction of RNA from heads of AC, SB, LB and PL, 12 embryos (two repetitions X six fish) at 178, 200 and 216 τs were used. Embryos were dechorionated and decapitated in front of the pectoral fin. RNA extraction and cDNA preparation were performed as described previously in 58. Similarly, RNA was extracted from a small piece (approximately 2 mm²) of skin, heart, liver, gill, spleen, intestine and liver from six adult AC-charr.

Analyses of RNA-seq data and mapping to Salmon EST contigs
As no *S. alpinus* genome is available and *de novo* assembly of the 36 bp reads yielded an excessive number of short contigs we chose to assess expression and genetic variation by mapping the reads to 59336 *S. salar* expressed sequence tag (EST) contigs from SalmonDB [63, downloaded 22. March 2012] and the Arctic charr mitochondrial genome [55, NC_000861].

To estimate expression, reads were aligned with RSEM version 1.1.18 with default parameters. RSEM distributes reads that map to multiple locations to the most likely contig, using expectation maximization[64]. The read counts for contigs with the same annotation were pooled because some genes were represented by more than one contig, and due to whole genome duplication salmonids have closely related paralogous genes[30,32]. Thus the expression tests are done on gene or paralog group level, instead of the contig level. In the remainder of the paper, the term 'gene' will have this broader meaning, some genes are represented by one contig and others by two or more (indicated in all relevant tables). This brought the number of genes considered down to 16851. Lastly, genes with fewer than 800 mapped reads in the entire dataset were excluded from the analyses, yielding a total of 10496 genes.

A generalized linear model (GLM) with morph and developmental time as explanatory variables was used to find genes with different expression levels between the two charr morphotypes (groups) using the edgeR-package in R[65].

$$Y = Morph + Time + Error$$

We could not test for interaction, as biological replicates were unavailable. To obtain further insight into the expression profiles of differently expressed genes, we preformed clustering analyses on log-transformed cpm-values (counts per million; cpm-function in edgeR). The values for each gene were scaled by mean and standard deviation, and the euclidean distance used for the hclust-function in R[66] with the default settings. We used the hypergeometric-test in goseq[67] to test for gene ontology enrichment. Since we pooled the

read-count from different contigs we could unfortunately not take gene length into account in those tests.

Assessment of differential expression with qPCR
We previously identified suitable reference genes to study Arctic charr development[58]. Here we examined the expression of several genes in whole charr embryos, embryonic heads and adult tissues. Primers were designed using the Primer3 tool[68] and checked for self-annealing and heterodimers according to the MIQE guidelines[69] (S1 Table). Primers for genes with several paralogs were designed for regions conserved among paralogs. For natterin, primers for the different paralogs were designed to match regions differing in sequence. Relative expression was calculated using the $2^{-\Delta\Delta Ct}$ method[70]. For the calculation of relative expression of genes in whole embryos, the geometric mean expression of three reference genes, *β-Actin*, *elongation factor 1α* and *Ubiquitin-conjugating enzyme E2 L3*, was used for normalization. For visual comparisons among samples, the normalized expression was presented as relative to the expression in AC at 161 τs (calibration sample). For the embryonic head samples *IF5A1* and *ACTB* were used as reference genes and a biological replicate of AC at 178 (τs) as the calibrator sample. Standard errors of relative expression were calculated from the standard errors (SE) of the ΔC_T-values with the formula $2^{-(\Delta\Delta Ct + SE)}$ = minimum fold expression and $2^{-(\Delta\Delta Ct - SE)}$ = maximum fold expression. The statistical analysis was performed using the ΔC_T-values with a two-way ANOVA with GLM function in R.

$$Y = Morph + Time + M \times T + Error$$

Normal distribution of residuals was confirmed for all data. For the study of expression in the embryonic head we followed a significant morph effect in the ANOVA with Tukey's post-hoc honest significant difference test, on relative expression ratios (ΔC_Ts).

Polymorphisms in the Arctic charr transcriptome
For analysis of genetic variation we mapped the reads to the salmon contigs, this time using the Burrows-Wheeler Aligner (BWA)[71] with a seed length of 25, allowing two mismatches. We re-mapped the reads, since BWA allows short indels (RSEM does not) but disregarding them leads to many false SNPs close to indels. To extract candidate polymorphic sites from the Arctic charr transcriptome we ran VarScan2[72] with minimum coverage of 50 reads and minimum minor allele frequency of 0.1 on reads mapped to each *S. salar* contig for all of the 8 timepoints and morph combinations. This was done separately for reads that mapped uniquely to one contig only (UNI) and reads that mapped to two or more contigs (REP). These SNP-candidates were further processed in R[66]. SNP-candidates at 90% frequency or higher in all samples were disregarded, as they reflect differences between Arctic charr and *S. salar* and are not the focus of this study. SNP-candidates with poor coverage in specific samples - i.e. coverage of five or fewer reads in three or four samples of each morph - were removed. As the SNP analysis was done on individual contigs, differences among paralogs appear in the data. However, since each sample is a pool of few individuals, it is very unlikely that we have the same frequency of true SNPs in the samples. This property was used to remove variants that are most likely due to expressed paralogs. Using Fisher exact tests to

evaluate differences between samples, only SNPs that were significantly different between samples with a $p < 0.05$ (with no multiple testing correction) were chosen for further examination. As equal cDNA input from individuals in sample cannot be assumed, due to expression differences among them and stochastic processes in sample preparation, read numbers were summed over the four samples from each morph for the comparison between them. A conservative approach was taken to look for differences between morphs. We focused on SNP-candidates that showed differences in frequency between morphs, without adjusting for multiple testing (Fisher exact test, $p > 5\%$). We extracted the most interesting candidates by filtering by frequency difference between the morphs (delta). SNP-candidates with the highest frequency difference (delta > 95%) were manually processed and redundant candidates removed. A similar approach was used to mine for polymorphisms in Arctic charr mtDNA (NC_000861), using *S. salar* mtDNA as the outgroup (NC_001960.1).

We wrote a python script to predict the impact of SNPs within the mRNA sequences. Polymorphisms were categorized according to their location (3'UTR, coding, 5'UTR), and those within the coding region into synonymous or non-synonymous.

Verification of candidate SNPs
We chose 12 candidate SNPs for verification (see below). The candidates were verified using a similar approach as previously[52]; first, we conducted genomic comparisons of the Salmon genome, ESTs and short contigs from the preliminary assembly of the Arctic charr transcriptome. This allowed us to infer the placement of the putative polymorphism in the locus, and design paralog specific primers for PCR (less than 1 kb amplicons) for verification of the 12 candidate SNPs (S2 Table). MJ tetrad machine was used for PCR and the program was 5 min. at 95°C, followed by 35 cycles of 30 sec. at 52°C, 1 min. at 72°C, 30 sec. at 95°C, ending with 12°C while waiting on the human. Each individual was genotyped by first amplifying the region of interest using PCR, followed by ExoSAP (Affymetrix), direct sequencing (BigDye) and finally run on an Applied Biosystems 3500xL Genetic Analyzer (Hitachi). Raw data was base-called using the Sequencing Analysis Software v5.4 with KBTMBasecaller v1.41 (Applied Biosystems). Ab1 files were run through Phred and Phrap and imported to Consed for visual editing of ambiguous bases and putative polymorphisms, and for trimming primers. The FASTA files were aligned with ClustalW online [73, http://www.ebi.ac.uk/Tools/msa/clustalw2/] and manually inspected in Genedoc[74]. All sequences where deposited to Genbank as popsets under the accession numbers KP019972-KP020026.

Comparative genomic analyses of sequence polymorphisms
Comparative genomics showed that several verified SNPs affected evolutionarily constrained parts of the mitochondrial genome. Two approaches were used: performing a BLAST search on salmon ESTs (May 2013) and retrieving multiZ alignments of vertebrates from the UCSC genome browser (in September 2013). This yielded several hundred sequences from related fish and other vertebrates. The list was reduced to 20 sequences for visualization, by keeping members of the major taxa but removing more closely related sequences,

aligned with ClustalW and manually adjusted in Genedoc. The species and genome versions used are; Human (*Homo sapiens*, hg19), Lamprey (*Petromyzon marinus*, petMar1), Fugu (*Takifugu rubripes*, fr2), Medaka (*Oryzias latipes*, oryLat2), Stickleback (*Gasterosteus aculeatus*, gasAcu1), Tetraodon (*Tetraodon nigroviridis*, tetNig2), Zebrafish (*Danio rerio*, danRer6). We also downloaded from NCBI the sequence of whole or partial mtDNA from several fish species; Brown trout (*Salmo trutta*, JQ390057 and AF148843), Broad whitefish (*Coregonus nasus*, JQ390058), Legless searsid (*Platytroctes apus*, AP004107), Pacific menhaden (*Ethmidium maculatum*, AP011602), Icefish (*Salanx ariakensis*, AP006231 and HM151535), Chain pickerel (*Esox niger*, AP013046) and Western Pacific roughy (*Hoplostethus japonicus*, AP002938). The three mitochondrial variants (numbered by the *S. alpinus* mtDNA - NC_000861) are; m1829G>A (CCACGTTGTGAAACCAAC[G/A]TCCGAAGGTGG ATTTAGCAGT), m3211T>C (CGTGCAGAAGCGGGCATAAG[T/C]ACATAAGACGAGAAGACCCT) and m3411C>T (CTCTAA-GCACCAGAATTT[C/T]TGACCAAAAATGATCCGGC).

Results
RNA sequencing characteristics
Each sample yielded good quality data, with sequencing depth from 49 to 58 million (average: 55 million) reads. To quantify the expression levels, the reads were aligned to a salmon EST-assembly[63]. Around 20% of the reads mapped uniquely to the EST data (S3 Table). A further 30% mapped to two or more contigs, probably representing paralogous genes, recent duplications or repeat-like elements within transcribed regions. A substantial fraction of the RNA-sequencing reads did not map to the contigs from *S. salar*. Analyses of those reads require an Arctic charr genome sequence or transcriptome assembly from longer and paired end reads.

Differential expression during Arctic charr development
For the expression analysis, ESTs were collapsed into 16851 "genes" or paralog groups (see the Methods for the broader meaning of gene). We only considered genes (total of 10496) with 800 or more reads mapped and tested for differential expression using the edgeR-package[65]. We detected considerable changes in the transcriptome during Arctic charr development (Figure 2a). The expression of 1603 and 2459 genes differed significantly between developmental timepoints at the 1% and 5% levels of false discovery rate (FDR), respectively (S1 file). The difference was most pronounced between pre-hatching (timepoints: 141, 163, 200 τs) and post hatching embryos (timepoint 433 τs), as more than 70% of the genes with FDR below 1% had higher expression in the latter (Figure 2a). According to Gene Ontology analyses, six separate GO categories are significant (below 10% FDR). The most drastic changes were seen in processes related to glycolysis (GO:0006096, FDR = 0.0009), were the expression of 19 out of 25 genes changed during this developmental period. The other five classes that were differentially expressed during charr development are: ion transport (GO:0006811, FDR = 0.027), blood coagulation (GO:0007596, FDR = 0.03), DNA repair (GO:0006281, FDR = 0.08) and two immune related categories (GO:0019882, FDR = 0.08, GO:0006955, FDR = 0.09). Those results probably reflect developmental changes and/or differences in the environment of embryos before and after hatching.
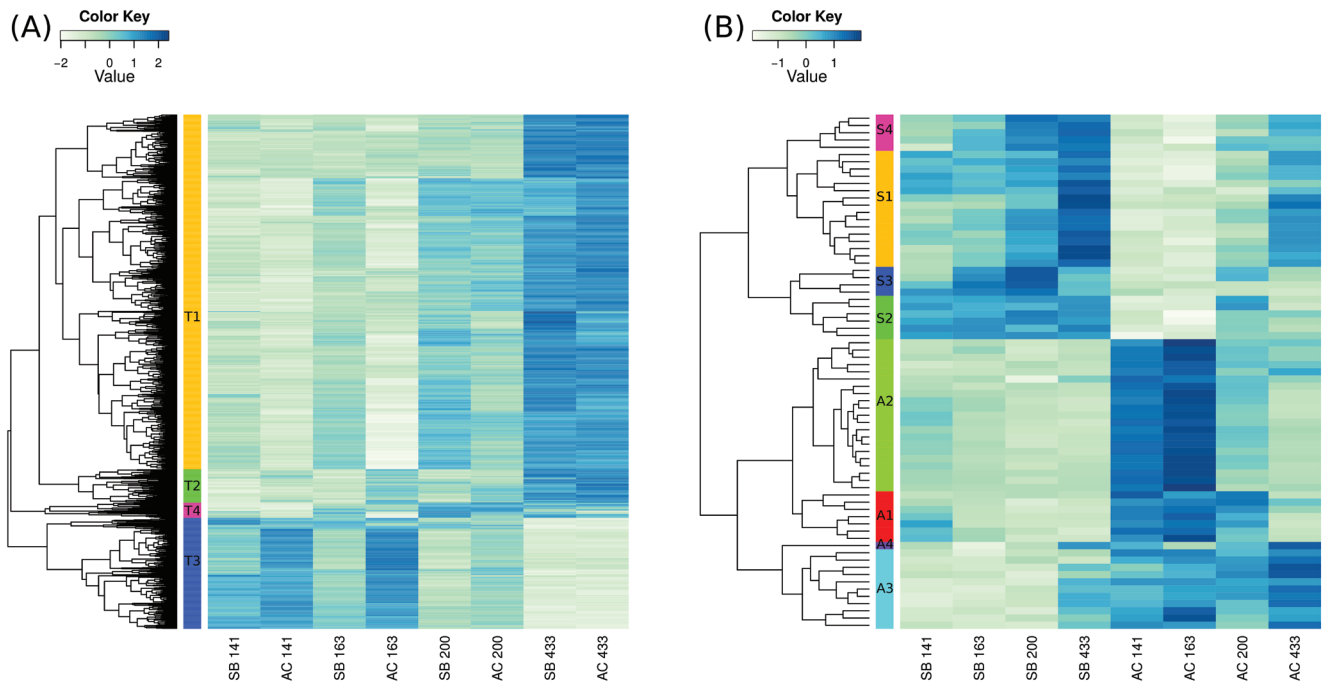
Figure 2. Heatmap of differentially expressed genes in the Arctic charr developmental transcriptome. Two morphs (SB and AC) are represented, at four timepoints. (**A**) The 1603 genes with expression difference among time points, here clustered into four groups . (**B**) The 71 genes differentially expressed between morphs, are clustered into 4 groups for each of the two morphs. High expression is indicated by blue and low expression by beige.

Differential expression between Arctic charr morphs

We were especially interested in genes showing expression differences between the two Arctic charr morphs as they might implicate pathways involved in the ecological divergence among charr populations. In the data 296 genes were differentially expressed (FDR < 5%) between the morphs (141 higher in SB and 152 higher in AC, S1 file). Among genes with higher expression in SB-charr two biological GO categories were enriched: blood coagulation (GO:0007596, p = 0.001) and proteolysis (GO:0006508, p = 0.002). Recall, expression of blood coagulation factors also differed between developmental stages (see above). In AC-charr, genes in three categories: respiratory electron transport chain (GO:0022904, p = 0.0006), ATP synthesis coupled electron transport (GO:0042773, p = 0.002) and neurotransmitter transport (GO:0006836, p = 0.009) have higher expression. The first two GO categories both relate to energy generation in mitochondria and could reflect higher expression of genes with mitochondrial functions in AC-charr.

At more stringent FDR (1%), 31 genes were higher expressed in SB and 40 genes higher in AC-charr (Figure 2b, Table 1 and Table 2). These genes have diverse functional annotations. The genes with higher expression in each morph were clustered into 4 groups, which aggregated genes of similar function. For instance SB cluster 3 has three immune related genes: *Complement factor D* (9), *H-2 class I histocompatibility antigen L-D alpha chain* (2) and *Sushi domain-containing protein 2* (4) and one gene with unknown function (Table 1). Note, however, that immune genes were not significantly enriched in the GO comparison of morphs.

The results suggest mitochondrial function and blood coagulation genes are differentially expressed between the morphs, but due to few samples used in the RNA-sequencing, qPCR verification was needed.

Validation of gene expression differences in whole embryos and paralog specific expression of *natterin* genes

The data highlights genes likely to differ in expression between embryos of SB and AC-charr. Of the nine genes subjected to qPCR analyses of whole embryos, five were confirmed to be differentially expressed between AC and SB at 161 or 200 τs (Figure 3, S4 Table and S2 file). Three genes, *Nattl*, *Alkaline phosphatase (Alp)* and *Lysozyme (Lyz)*, had significantly higher expression in SB. The other two, *Keratin-associated protein 4–3 (Krtap4–3)* and *Poly polymerase 6 (Parp6)* had higher expression in AC embryos (Figure 3, S4 Table). No morph and time interaction was detected for any of the genes.

As some of the genes are represented by different contigs or even paralogs, we set out to disentangle the expression of one paralog group in detail. The qPCR primers used above matched conserved gene regions and thus estimate the combined expression of several paralogs. We chose to measure the expression of three different *natterin* paralogs (*nattl1, 2 and 3*), in part because this understudied gene was first characterized as a toxin produced by a tropical fish[75,76]. We studied *nattl* expression in several developmental stages in AC-, SB- and PL-charr as well as in selected tissues of adult AC-charr. The expression level of the three paralogs differed between

Table 1. Differentially expressed genes, with higher expression in the SB morph from Lake Thingvallavatn.

NR	Unigene.Description	NR.contigs	logFC	logCPM	FDR	Cluster
3766	Histone H3 embryonic	1	8.71	2.74	7.80E-035	S-1
5103	Natterin-like protein	6	2.75	7.12	7.76E-007	S-2
356	A7J6M9 Putative uncharacterized protein n175R	1	2.33	4.66	3.30E-006	S-1
6697	Q1KY05 Main olfactory receptor-like protein	5	3.12	6.92	9.96E-005	S-1
8151	Sushi domain-containing protein 2	4	2.20	5.55	0.0001	S-3
1682	Carcinoembryonic antigen-related cell adhesion molecule 1	3	2.55	3.83	0.0002	S-1
6228	Protein FAM98A	2	1.96	4.76	0.0003	S-1
7531	STAM-binding protein-like	2	2.07	2.62	0.0005	S-1
6712	Q1M160 Myc-regulated DEAD box protein	1	1.67	3.23	0.0009	S-1
2300	Cytosolic sulfotransferase 3	3	1.73	2.13	0.0009	S-1
2063	Complement factor D	7	1.79	6.42	0.0016	S-3
3326	Galectin-3-binding protein A	4	1.79	3.85	0.0017	S-4
3169	Flocculation protein FLO11	2	1.86	4.05	0.0017	S-1
1203	B5XDY0 H-2 class I histocompatibility antigen L-D alpha chain	2	1.70	2.12	0.0028	S-3
9183	UPI000065D844 UPI000065D844 related cluster	2	1.97	5.55	0.0028	S-1
2909	Epidermis-type lipoxygenase 3	4	1.68	4.84	0.0029	S-1
4884	Myeloperoxidase	4	2.20	6.78	0.0029	S-1
10003	Uridine phosphorylase 1	4	1.51	3.00	0.0047	S-1
2513	Desmoglein-1-alpha	1	1.59	2.80	0.0054	S-2
377	A7SJA8 Predicted protein (Fragment)	1	1.73	2.50	0.0055	S-3
9204	UPI00006A2900 UPI00006A2900 related cluster	2	6.38	3.26	0.0064	S-1
9642	UPI00017B1B0F UPI00017B1B0F related cluster	1	2.00	1.92	0.0064	S-2
1965	Coiled-coil domain-containing protein 136	2	2.15	2.32	0.0064	S-2
9260	UPI0000F1D4BA PREDICTED: hypothetical protein	1	1.80	2.41	0.0065	S-2
738	Adseverin	8	1.58	5.51	0.0073	S-1
9678	UPI00017B4479 UPI00017B4479 related cluster	1	2.18	1.97	0.0074	S-4
8339	Testin	4	1.50	4.93	0.0080	S-2
6840	Q4SNH3 Chromosome 8 SCAF14543 whole genome shotgun sequence	1	1.42	4.00	0.0080	S-1
1668	Carbohydrate sulfotransferase 6	1	2.09	2.08	0.0090	S-4
8341	Testisin	2	2.01	2.76	0.0090	S-4
6373	Protein asteroid homolog 1	6	1.29	4.24	0.0090	S-4

logFC – log Fold Change

logCPM – log Counts Per Million

FDR – False Discovery Rate

The cluster numbering corresponds to Figure 1.

Table 2. Differentially expressed genes, with higher expression in the AC morph.

NR	Unigene.Description	NR.contigs	logFC	logCPM	FDR	Cluster
3465	Glutathione S-transferase P 1	1	-8.35	2.45	1.12E-019	A-2
2475	Dehydrogenase/reductase SDR family member 7	2	-4.88	2.15	9.67E-014	A-3
6945	Q6NWE8 Sb:cb283 protein (Fragment)	3	-6.08	3.02	2.15E-013	A-2
399	A8DW32 Predicted protein	1	-5.32	6.38	4.27E-010	A-1
9682	UPI00017B4B48 UPI00017B4B48 related cluster	2	-3.70	2.81	2.61E-008	A-2
9817	Uncharacterized protein ART2	5	-12.63	6.89	8.23E-008	A-2
6724	Q2L0Z2 Putative ATP-dependent RNA helicase	1	-3.41	1.89	1.88E-007	A-2
1197	B5XD10 Vacuolar proton pump subunit G 1	1	-4.30	2.10	1.84E-006	A-2
5325	Nucleoside diphosphate kinase B	1	-9.85	7.63	2.51E-006	A-1
9205	UPI0000D5B923 PREDICTED: myelin basic protein isoform 1	3	-2.49	3.45	9.18E-006	A-3
6377	Protein broad-minded	1	-2.11	2.74	4.75E-005	A-1
5711	Pistil-specific extensin-like protein	1	-2.16	2.60	0.0002	A-3
3203	Formin-like protein 20	7	-1.98	1.95	0.0002	A-3
9315	UPI0000F2EC69 PREDICTED: hypothetical protein	2	-5.60	4.57	0.0005	A-2
363	A7RFV0 Predicted protein (Fragment)	2	-1.74	4.96	0.0010	A-3
6937	Q6AZT1 MGC81677 protein	3	-2.06	3.81	0.0014	A-2
3756	Histone H1	3	-2.26	4.54	0.0017	A-2
1133	B5DGN9 Creatine kinase-1	7	-4.72	5.50	0.0017	A-3
309	A1IMH7 CD80-like protein	12	-1.94	4.29	0.0017	A-2
7651	Serine protease ami	2	-1.54	5.90	0.0017	A-3
9935	Uncharacterized protein C7orf63 homolog	1	-1.87	1.91	0.0025	A-2
5219	Nostrin	2	-2.55	3.38	0.0029	A-2
1855	Chondroitin sulfate N-acetylgalactosaminyltransferase 2	5	-2.56	6.14	0.0034	A-1
10203	Xylose isomerase	6	-1.55	2.43	0.0035	A-3
2249	Cytochrome c oxidase subunit 3	11	-1.78	11.15	0.0035	A-3
180	40S ribosomal protein S3-B	2	-5.31	8.67	0.0050	A-1
1227	B6NBL3 Putative uncharacterized protein	3	-1.59	2.95	0.0050	A-2
5055	NADH-ubiquinone oxidoreductase chain 6	2	-1.48	2.65	0.0061	A-3
4634	Metallothionein A	1	-3.33	5.44	0.0064	A-2
342	A5C0J4 Putative uncharacterized protein	2	-2.58	2.47	0.0064	A-2
9698	UPI00019258B4 PREDICTED: similar to epithelial cell transforming sequence 2 oncogene protein partial	1	-2.06	2.94	0.0064	A-2
5878	Pro-opiomelanocortin B	1	-2.04	5.60	0.0065	A-2
2248	Cytochrome c oxidase subunit 2	9	-2.21	9.83	0.0074	A-2
1246	B8JI87 Novel protein similar to vertebrate collagen type VI alpha 3 (COL6A3) (Fragment)	1	-1.69	3.16	0.0080	A-3
7994	Sperm-associated antigen 5	1	-2.07	3.71	0.0080	A-2
9515	UPI000175F90F PREDICTED: similar to pleckstrin homology domain containing family A member 7	1	-2.00	1.87	0.0090	A-2
1124	B5DDZ4 Acta1 protein	1	-1.52	2.62	0.0090	A-2
1127	B5DG94 2-peptidylprolyl isomerase A	2	-2.56	5.67	0.0090	A-1
9175	UPI000054A3C0 PREDICTED: apolipoprotein B	3	-1.32	4.09	0.0090	A-4
9671	UPI00017B3C62 UPI00017B3C62 related cluster	1	-1.51	1.92	0.0096	A-1

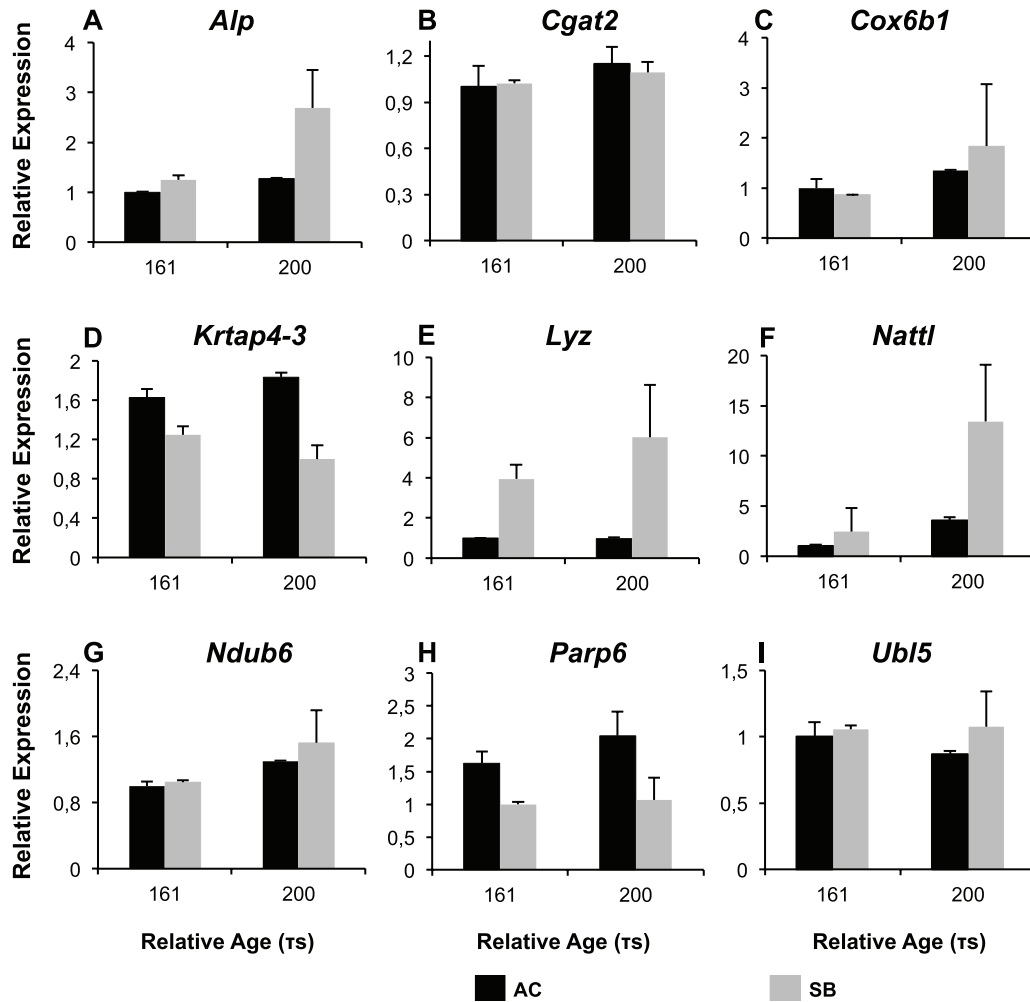For column header explanation, see footer of Table 1.

Figure 3. qPCR validation of candidates from transcriptome in whole embryos of Arctic charr. Relative expression of 9 genes (**A**–**I**) analysed by qPCR in the small benthic (SB) charr from Lake Thingvallavatn and aquaculture (AC) charr at two different developmental timepoints (161 and 200 τs). 5 genes were differentially expressed between the two morphs (*Alp, Krtap4-3, Lyz, Nattl, Parp6*), while 4 further genes did not show significant expression differences between morphs (*Cgat2, Cox6B1, Ndub6, Ubl5*), see Table S3. Error bars represent standard deviation calculated from two biological replicates.

morphs and timepoints (Figure 4 and S5 Table). Overall *nattl2* had the highest expression in all morphs. The *nattl1* had higher expression in embryos of PL-charr than in AC- and SB-charr, while *nattl2* and *nattl3* were more expressed in SB-embryos.

In order to evaluate the hypothesis that *nattl* genes have immune-related functions we studied expression in adult tissues (in AC-charr). The *nattl* expression was highest in the gills, followed by expression in kidney, skin and spleen. Low expression levels were detected in liver, intestine and heart (S1 Figure and S5 Table). The three *nattl* paralogs followed different patterns, whilst each of them showed significant expression differences among tissues. *Nattl1* was mainly expressed in spleen and kidney, while *nattl2* showed a significantly higher expression in skin, liver and in gills. Similarly, the relative expression of *nattl3* was highest in the gills and skin. This indicates that the three *nattl* paralogs are expressed in a tissue specific manner, and also differently during the development of the three charr morphs studied here.

Expression differences in the developing heads of benthic and limnetic charr morphs

To study the craniofacial divergence between sympatric Arctic charr morphs we used qPCR to study 8 genes with expression difference in the RNA-seq data (all higher in SB). We focused on genes with known craniofacial expression in zebrafish development[77] and compared two benthic (SB, LB) and two limnetic charr (AC, PL). We analyzed heads at three time-points (178, 200 and 218 τs) as this period overlaps with early stages of craniofacial skeletal formation in Arctic charr[78,79]. The qPCR confirmed the higher expression of seven out of these eight genes in the head of benthic charr compared to limnetic charr (Figure 5, S2 Figure and S3 file). These seven genes are *Claudin 4 (Cldn4), adseverin (Scin), Junction plakoglobin (Jup), Lipolysis stimulated lipoprotein receptor (Lsr), Major vault protein (Mvp), Transforming growth factor beta receptor II (Tgfbr2)* and *Vitamin D receptor a (Vdra)*. The eighth gene, *Retinoic acid receptor gamma-A (Rarg)* gave a small but significant
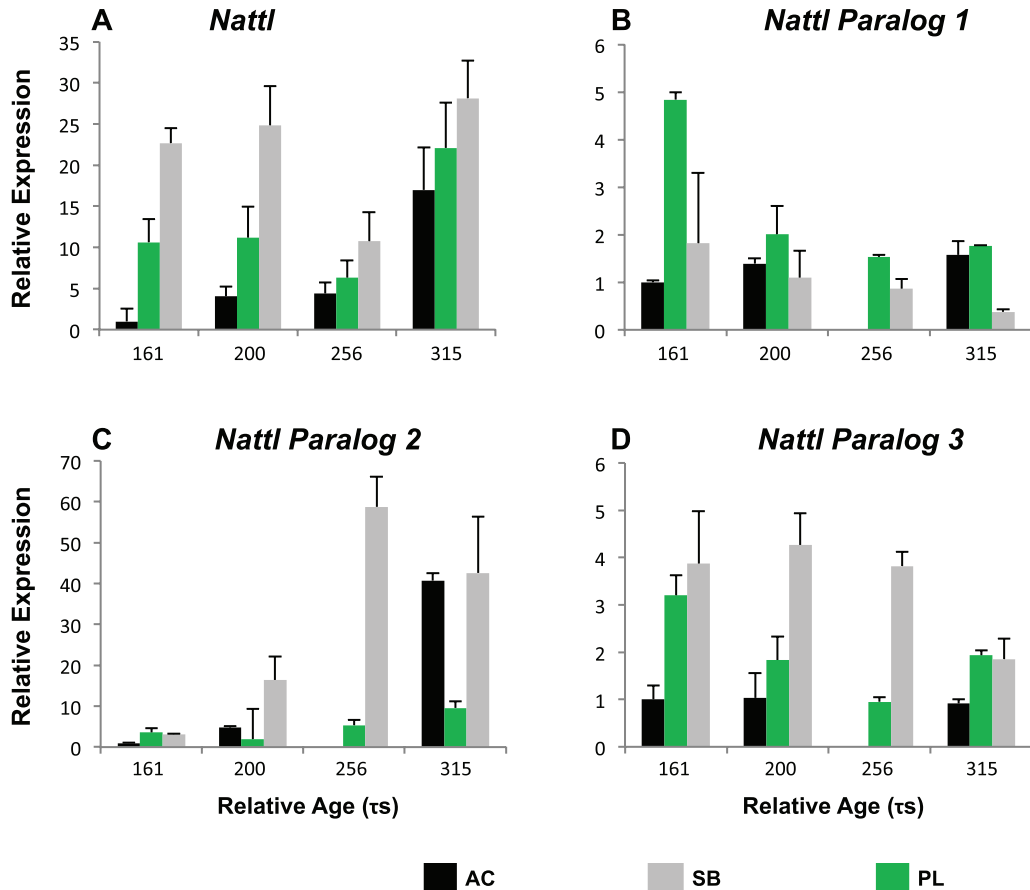
Figure 4. Relative expression of *Nattl* and its three paralogs during charr development in different morphs. The expression is graphed for different morphs (SB, AC and PL) at four developmental timepoints (161, 200, 256 & 315 τs, relative to AC-charr at timepoint 161. **A**) General {*Nattl*} expression along charr development. **B**–**D**) Expression of *Nattl paralogs* 1–3. ANOVA showing the variation among morphs is summarized in Table S4.

Gene	Morph (p-value)	HSD morph				Time (p-value)	HSD time (Ts)			M x T (p-value)
Cldn4	5.30e-07	AC	PL	SB	LB	4.77e-05	178	200	216	2.64e-03
Jup	7.18e-05	AC	PL	SB	LB	8.18e-04	178	200	216	0.754
Lsr	1.6e-06	AC	PL	SB	LB	2.10e-04	178	200	216	0.462
Mvp	1.25e-05	AC	PL	SB	LB	1.91e-03	178	200	216	0.94
Rarg	0.031	AC	PL	SB	LB	1.37e-03	178	200	216	0.321
Scin	2.43e-07	AC	PL	SB	LB	0.26		NS		3.53e-04
Tgfbr2	1.85e-05	AC	PL	SB	LB	8.48e-07	178	200	216	0.021
Vdra	2.20e-04	AC	PL	SB	LB	3.68e-06	178	200	216	0.285

Figure 5. Expression differences of craniofacial candidate genes in developing head of Arctic charr morphs. Relative expression ratios, calculated from the qPCR data, were subjected to an ANOVA to test the expression differences amongst four charr groups and three close time points (τs). The underlined gene names reflect significant difference between SB and AC-charr. A post hoc Tukey's test (HSD) was performed to determine the effects of morphs, time and morph-time interaction (M X T). White boxes represent low expression, while black boxes represent high expression. The shading represents significant different expression between the samples (p = 0.05, NS = not significant).

response in the head, but the effects were reversed, i.e. the expression was higher in AC. The expression difference of the seven genes was, in almost all cases, consistent over the three time-points studied (See S2 Figure).

Analyses of polymorphism in Arctic charr transcriptome

The RNA-seq data also revealed segregating variations with large frequency differences between charr morphs. To uncover candidate SNPs we mapped the reads to all of the *S. salar* EST-contigs. Filtering on coverage yielded 165,790 candidate SNPs (Table 3); of those 66.569 came from reads that mapped uniquely and 57.009 candidate SNPs from reads that mapped to more than one contig; with limited overlap between lists. Assuming that the expression of paralogous genes is stable, then differences among paralogs appear as SNPs at similar frequency in all samples. By requiring variant frequency differences (p < 0.05, uncorrected) between samples we reduced the list of candidates by two thirds, yielding over 20.000 candidate SNPs. Note, that as cDNA from charr families was sequenced (not a population sample), estimates of SNP frequencies are imprecise. To err on the side of caution, we chose SNP candidates with 50% or higher frequency difference between morphs for further study. The candidate SNPs were also summarized by frequency of the derived allele, in reference to the *S. salar* sequence. This gave 672 and 872 SNPs at higher frequency, in AC-charr and SB-charr, respectively. The uniquely and multiply mapped reads, revealed approximately similar numbers of candidate SNPs. Gene ontology analysis showed that for derived SNPs in SB, there was an excess of variants in genes related to translation, both as a broad category and specific subgroups (S6 Table). There was also enrichment of SNPs in genes related to DNA-mediated transposition, DNA integration, DNA replication and oxidation-reduction process. No GO categories were enriched for high frequency derived

Table 3. Candidate SNPs in the Arctic charr transcriptome, filtered by coverage, difference between sample and morphs and frequency difference between morphs.
For Delta > 0.95 we show the number of SNP-candidates before the redundant ones were removed.

SNP-candidates	Morph	Uni	Rep	Total
Total		96231	74341	165790
Filter coverage		66569	57009	113776
Diff. Bwn. samples		21417	22252	42869
Diff. Bwn. morphs		11385	12953	23974
Delta > 0.5	AC	396	285	672
Delta > 0.5	SB	526	353	872
Delta > 0.75	AC	95	68	159
Delta > 0.75	SB	155	95	248
Delta > 0.95	AC	17	13	30
Delta > 0.95	SB	29	4	33

SNP-candidates are found by mapping to *S. salar* ESTs. From UNIquely or REPeatedly mapped RNA-reads. Delta: Differences in allele frequency between morphs, categorized by which morph had the higher derived allele frequency.

SNPs in AC. Furthermore, functional effects of the candidate SNPs (UTR, synonymous and non-synonymous) were predicted. The distribution among those categories did not differ between variants detected by uniquely or repeatedly mapped reads, $\chi^2_{[3]} = 2.59$, $p = 0.46$ (S7 Table).

A total of 60 candidate SNPs are nearly fixed in one morph, with frequency difference between morphs above 95% (after manual inspection of contigs and SNP position three candidates were removed since they represented the same SNP). Of these "fixed" SNPs 46 came from uniquely mapped reads and 14 from reads that mapped more than twice (Table 4 and Table 5). For the SNPs from uniquely mapped reads, 17 are fixed in AC-charr and 29 in SB-charr. The few genes with two or more polymorphic sites were; *Keratin type II cytoskeletal 3 (KRT3)*, *Cysteine sulfinic acid decarboxylase (CSAD)* and *DNA-directed RNA polymerase I subunit RPA12 (RPA12)* with 5, 5 and 2 SNPs respectively. *KRT3* and *CSAD* had significant differentiation in both SB and AC. Similarly, 14 SNPs with large differentiation between morphs were predicted from reads that mapped on two or more contigs (Table 5). Of these, we found two variants in the mitochondrial *60S ribosomal protein L36 (RPL36)* and variants in 4 other mitochondrial genes (*28S ribosomal protein S18a mitochondrial (MRPS18A)*, *Apoptosis-inducing factor 1 mitochondrial (AIFM1)*, *Isocitrate dehydrogenase [NADP] mitochondrial (acIDH1)* and *Protein S100-A1 (S100A1)*), all at higher frequency in AC-charr. PCR and Sanger sequencing was used to confirm SNPs in *DNA2-like helicase (DNA2)*, a gene with nuclear and mitochondrial function, and two other genes *Uroporphyrinogen decarboxylase (UROD)*, and *Mid1-interacting protein 1-like (MID1IP1)* (S2 Table). The candidate variant *Eukaryotic translation initiation factor 4 gamma 2 (EIF4G2)* was not substantiated by the PCR/sequencing.

Polymorphism and expression of Arctic charr mtDNA

Considering the enrichment of differentially expressed genes related to mitochondrial energy metabolism (above), and high frequency candidate SNPs in several genes with mitochondrial function in AC-charr we decided to study the mitochondrial transcriptome further. The charr studied here reflect metabolic extremes, the aquaculture charr was bred for growth while the small benthic morph is thought to have experienced natural selection for slow metabolism and retarded growth[45,80]. Although mRNA preparation protocols were used for generating cDNA for the RNA-sequencing, a substantial number of reads came from non-polyadenylated sequences. By mapping the reads to mtDNA sequence of Arctic charr we could estimate expression and infer polymorphism both in genes and intergenic regions. There was a clear difference in sequencing coverage, with more than twice as many reads mapped from the AC- compared to SB-charr (mean fold difference 2.27, Wilcoxon test, p < 0.0004). Note, as only two types of fish are compared, it is impossible to determine the polarity of expression divergence.

The mapped RNA-reads were used to identify polymorphism and divergence in the entire mitochondrial chromosome. The polymorphisms were found by mapping to mtDNA from a Canadian *S. alpinus*[55], but ancestral vs. derived status inferred by comparison to *S. salar* mtDNA. Bioinformatics revealed 82 candidate sites,

including 35 that represent divergence between Icelandic and Canadian charr. A total of 20 candidate SNPs had high (more than 50%) frequency difference between SB and AC-charr (Figure 6). There was no bias in the distribution of derived SNPs, 11 on the AC branch and 9 in SB. Note, the frequency distribution is most irregular as we sequenced embryos of related individuals (see Materials and Methods), not a population sample. The divergence between Iceland and Canada is particularly little in the 12s and 16s ribosomal RNA genes. Curiously in those genes were two SNPs differing strongly in frequency between morphs (Figure 6). To confirm and better estimate the frequency of variants in the ribosomal genes, we PCR amplified and sequenced two ~550 bp regions in the rRNA genes, comparing three morphs (PL, LB and SB) from Lake Thingvallavatn (Figure 7A, C and E, S2 Table). The 12s polymorphism (m1829G>A) differed significantly between the morphs ($\chi^2_{[2]}$ = 8.6, p = 0.014), and was at highest frequency in the SB (0% in PL, 12.5% in LB and 75% in SB). Similarly m3411C>T in the 16s was enriched in SB (62.5%) but found at lower frequency in

PL (0%) and LB (12.5%) (it differed significantly between morphs, $\chi^2_{[2]}$ = 9.3333, p = 0.009). The Sanger sequencing also revealed three other polymorphisms in the amplified region, not seen in the transcriptome. Among those m3211T>C in the 16s gene was at 75% frequency in LB, but not found in the other morphs ($\chi^2_{[2]}$ = 19.76, p < 0.0001).

In order to gauge the potential functionality of those variants we aligned the rRNA genes from nearly hundred fishes and several vertebrates. The position affected by m1829G>A and m3211T>C, in the 12s and 16s rRNAs, are not well conserved in fishes or vertebrates (Figure 7B and Figure 7D). However m3411C>T, in the 16s rRNA, alters a position that is nearly invariant in 100 fish genomes (Figure 7F). The only exception is Pacific menhaden, which curiously also has T in this position. This region could not be aligned properly in other vertebrates. Thus m3411C>T alters a conserved position, but probably not very drastically as the introduced allele is tolerated in another fish species.

Table 4. SNP candidates from uniquely mapped reads.

(a) Higher frequency in AC morph								
Contig	Annotation	Pos	Ref	Var	Freq-SB	Freq-AC	Effect	
SS2U026955	Keratin type II cytoskeletal 3	300	A	T	0.000	0.984	synonymous	
SS2U026955	Keratin type II cytoskeletal 3	309	G	A	0.000	0.996	synonymous	
SS2U033960	Cysteine sulfinic acid decarboxylase	192	C	G	0.000	1.000	5prime	
SS2U033960	Cysteine sulfinic acid decarboxylase	416	G	T	0.000	0.961	G to V	
SS2U033960	Cysteine sulfinic acid decarboxylase	945	C	A	0.004	0.956	synonymous	
SS2U043396	Eukaryotic translation initiation factor 2-alpha kinase 1	134	A	G	0.000	1.000	5prime	
SS2U043886	Transcription cofactor HES-6	1308	T	C	0.000	1.000	5prime	
SS2U044339	Intraflagellar transport protein 52 homolog	479	T	C	0.021	1.000	D to G	
SS2U045168	Putative Peptide prediction	1275	G	A	0.000	1.000	3prime	
SS2U045328	E3 ubiquitin-protein ligase DTX3L	388	G	A	0.000	0.977	synonymous	
SS2U045990	Low-density lipoprotein receptor-related protein 1	135	T	C	0.000	0.969	synonymous	
SS2U048125[a]	Transmembrane protein 131-like	480	G	A	0.000	1.000	synonymous	
SS2U052747	Uridine 5'-monophosphate synthase	914	G	A	0.000	0.951	synonymous	
SS2U054542	Mediator of RNA polymerase II transcription subunit 20	474	C	T	0.027	0.995	synonymous	
SS2U056193	SUMO-conjugating enzyme UBC9	96	A	T	0.000	1.000	3prime	
SS2U057101	ETS domain-containing protein Elk-3	440	C	G	0.000	1.000	3prime	
SS2U058860	Voltage-dependent anion-selective channel protein 2	681	G	T	0.000	1.000	3prime	

(b) Higher frequency in SB morph								
Contig	**Annotation**	**Pos**	**Ref**	**Var**	**Freq-SB**	**Freq-AC**	**Effect**	
SS2U000399	Insulin-like growth factor-binding protein 7	598	C	A	1.000	0.000	3prime
SS2U004484	Titin	387	G	A	0.990	0.010	synonymous
SS2U026826	L-asparaginase	363	C	T	1.000	0.000	H to Y
SS2U026955	Keratin type II cytoskeletal 3	116	C	A	0.996	0.031	T to N
SS2U026955	Keratin type II cytoskeletal 3	264	C	T	0.970	0.008	synonymous
SS2U026955	Keratin type II cytoskeletal 3	317	C	T	1.000	0.002	T to M
SS2U033960	Cysteine sulfinic acid decarboxylase	363	C	T	1.000	0.025	5prime
SS2U033960	Cysteine sulfinic acid decarboxylase	387	C	T	1.000	0.030	synonymous
SS2U033960	Cysteine sulfinic acid decarboxylase	657	T	C	0.990	0.031	synonymous
SS2U034322	Cyclin-C	1094	A	G	1.000	0.000	3prime
SS2U034431	Dolichyl-diphosphooligosaccharide–protein glycosyltransferase subunit 2	436	G	A	0.992	0.000	G to S
SS2U036025	Nuclear receptor coactivator 4	36	G	A	1.000	0.043	5prime
SS2U040590	Glutamyl-tRNA(Gln) amidotransferase subunit A homolog	478	G	A	0.972	0.000	synonymous
SS2U045606	Superkiller viralicidic activity 2-like 2	500	C	T	1.000	0.000	synonymous
SS2U047816	Squalene synthase	1139	G	A	1.000	0.029	synonymous
SS2U048063	Lysine-specific demethylase NO66	669	C	T	1.000	0.000	synonymous
SS2U050394	UPF0542 protein C5orf43 homolog	596	G	A	1.000	0.000	synonymous
SS2U050880[a]	Transmembrane protein 131-like	901	C	T	1.000	0.000	A to V
SS2U052076	Eukaryotic translation initiation factor 3 subunit A	824	C	T	1.000	0.031	synonymous
SS2U053417	RNA polymerase-associated protein LEO1	454	G	A	1.000	0.049	synonymous
SS2U054333	Scaffold attachment factor B2	382	G	A	0.999	0.000	V to M
SS2U054705	Cell division protein kinase 4	122	A	G	0.971	0.000	3prime
SS2U054965	DNA-directed RNA polymerase I subunit RPA12	106	G	A	1.000	0.000	5prime
SS2U054965	DNA-directed RNA polymerase I subunit RPA12	411	T	G	1.000	0.000	synonymous
SS2U055120	Chromatin modification-related protein MEAF6	350	A	C	1.000	0.000	H to P
SS2U055153	Complexin-1	1191	C	A	1.000	0.031	3prime
SS2U057635	Mitogen-activated protein kinase 14B	1370	A	T	1.000	0.026	3prime
SS2U058169	Transmembrane protein 50A	1214	C	G	0.973	0.000	3prime
SS2U058802	Signal recognition particle 54 kDa protein	607	T	A	0.969	0.000	C to S

[a] Those genes are distinct paralogs

Table 5. **SNP candidates with significant difference frequency between AC and SB morphs, from reads that mapped to two or more contigs.**

| Contig	Annotation	Pos	Ref	Var	Freq-SB	Freq-AC	Effect
SS2U004839	Actin alpha sarcomeric/cardiac	550	A	C	0.015	0.999	3prime
SS2U021298	28S ribosomal protein S18a mitochondrial	462	A	C	0.000	1.000	synonymous
SS2U041264	Apoptosis-inducing factor 1 mitochondrial	341	C	T	0.000	0.952	synonymous
SS2U054211[a]	Cytoplasmic dynein 1 intermediate chain 2	136	T	C	0.018	0.974	synonymous
SS2U054362[a]	Q08CA8 Dynein cytoplasmic 1 intermediate chain 2	945	A	G	0.000	1.000	synonymous
SS2U055923	Bystin	1623	A	C	0.000	0.983	3prime
SS2U058758	Protein S100-A1	253	C	T	0.000	0.984	synonymous
SS2U059000	Isocitrate dehydrogenase [NADP] mitochondrial	1654	T	C	0.000	0.975	3prime
SS2U059146	60S ribosomal protein L36	263	T	G	0.009	1.000	synonymous
SS2U059146	60S ribosomal protein L36	470	A	C	0.009	1.000	synonymous
SS2U036667	Heterogeneous nuclear ribonucleoprotein K	813	C	T	1.000	0.022	5prime
SS2U042873	RNA polymerase-associated protein LEO1	460	G	A	1.000	0.000	synonymous
SS2U058455	Adenylosuccinate lyase	1616	C	T	1.000	0.000	3prime
SS2U058906	Mid1-interacting protein 1-like	350	G	T	0.985	0.000	E to D

[a] Those genes are distinct paralogs
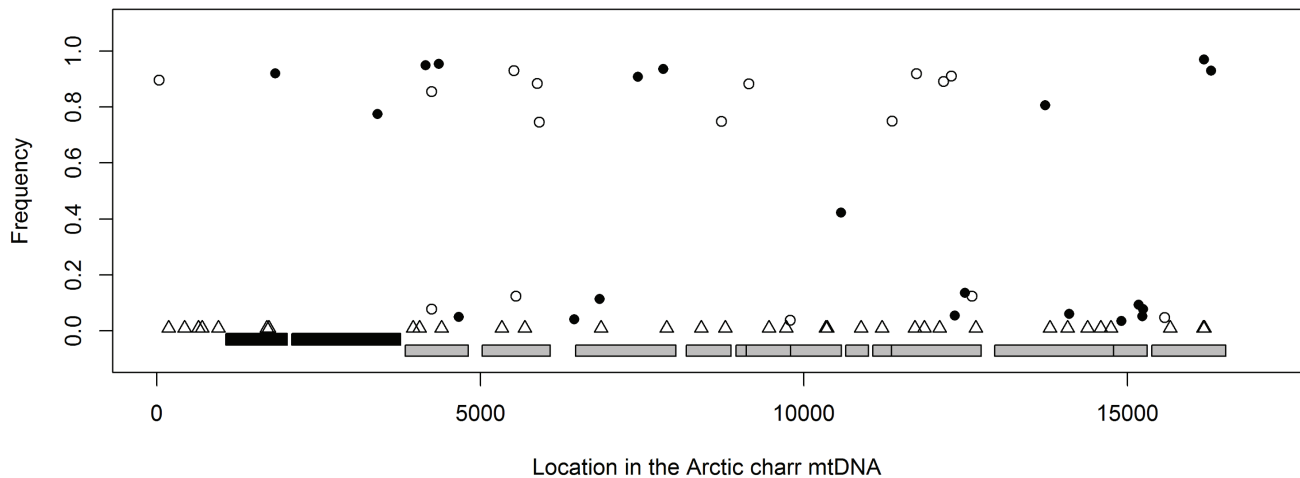
Figure 6. Genetic divergence in the mtDNA between SB- and AC-charr. The frequency differences between morphs of candidate SNPs, estimated from the RNA-sequencing, graphed along the mtDNA chromosome. The SNPs indicate whether the derived allele is of higher frequency in SB (black dots) or AC (open circles). Sites of divergence between the Icelandic stocks and the Canadian reference sequence are indicated by triangles. The two black boxes represent the 12s (left) and 16s (right) rRNA genes, and gray boxes the 14 coding sequences.
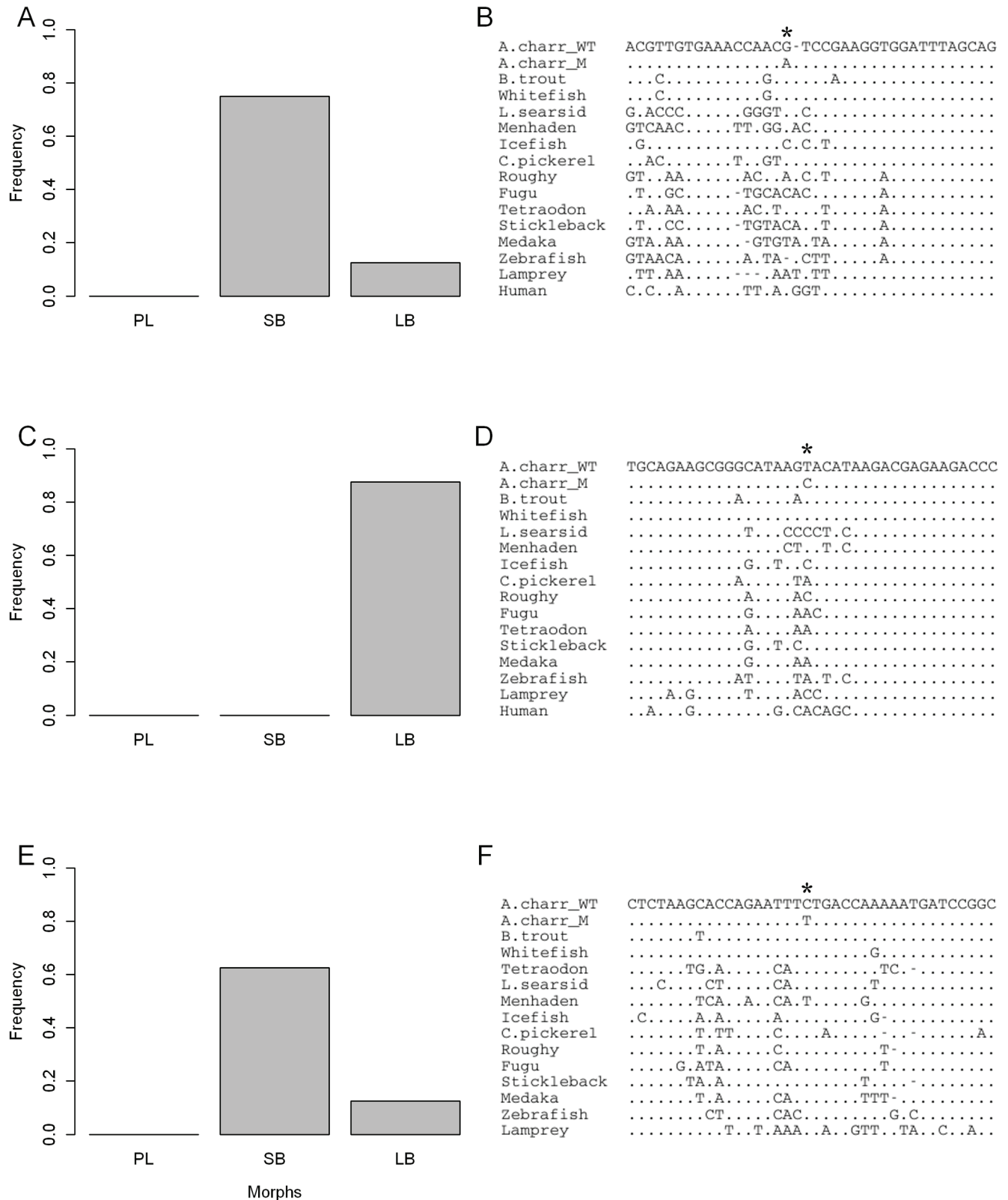
Figure 7. Comparative genomics and population genetic differentiation in Arctic charr at 3 mtDNA locations. Aligned are several fish genomes, with Lamprey or humans as outgroups, reflecting a 38 bp window around each of the 3 positions (*). **A**, **C**, **E**) Frequency of each of those variants in three Arctic charr populations from Lake Thingvallavatn (PL, LB and SB). A total of 8 individuals were genotyped from each morph, see methods. **B**) Alignment of variant m1829G>A in the 12s rRNA gene in fishes, using humans as an outgroup. **D**) Similar alignment of a 16s variant, m3211T>C and **F**) alignment of variant m3411C>T in the 16s rRNA gene.

Discussion

We are interested in how predictable evolution is a the molecular level and if there certain principles influence the rewiring of developmental and regulatory systems during evolution[81]. One way to study this is to identify genetic and developmental effects affecting key traits in species or populations which exhibit parallel evolution. The objective of this study were to get a handle on genetic and molecular systems that associate with benthic morphology in charr by mainly focusing on the small benthic morph in Lake Thingvallavatn, Iceland. To this end we performed transcriptome analysis contrasting the development of embryos from SB-charr and aquaculture charr.

Developmental transcriptome of Arctic charr morphs

As no reference genome is available for Arctic charr, we mapped reads to *S. salar* EST-contigs[63] in order to estimate expression and identify candidate genetic polymorphisms. As many of the contigs

are short or have overlapping annotations, we collapsed genes into paralogous genes when appropriate for the expression analysis. The main advantage of this approach was the reduction of the number of statistical tests (and hence an increase in statistical power). The downside is that paralog-specific expression patterns are masked, as our qPCR results of the *natterin like* gene family show (Figure 3 and S1 Figure). Recent rainbow trout data shows most paralogs from the latest whole genome duplication event retain the same expression pattern[32] indicating that this scenario is probably uncommon; hence it is of considerable interest when two paralogs show distinct expression patterns[82]. In their analysis of the Arctic charr gill transcriptome, Norman *et al.* (2014)[16] also used Illumina sequencing technology to evaluate expression. Their reads were longer (2x100 bp) than in this study (36 bp) enabling them to assemble contigs. They did not consider the paralogs in their approach and merged contigs based on sequence identity. Thus the complexity of Arctic charr transcriptome still remains a mystery that advances in sequencing technology, assembly algorithms and genome sequencing of this species could aid in revealing.

Our data reflects differential deployment of several gene classes during Arctic charr development. Studies in salmonids and other fish have demonstrated large changes in expression during early development, including coordinated changes in many cellular and developmental systems[12,19,83–85]. Several blood coagulation factors genes showed significant changes during charr development, and were also more highly expressed in the SB-charr. This might reflect differences in the rate of development of blood composition, or tissue composition, in the two morphs. Our main interest is on the derived and repeatedly evolved small benthic charr. For this study we chose AC-charr as a point of reference for several reasons, i) it has limnetic like head morphology, ii) availability, and iii) because we wanted a strong contrast in this first survey of charr developmental diversity. The AC-charr proved a useful, as the data presented here has already revealed differential expression of several developmental genes and regulators with differential expression between benthic and limnetic charr[58,59]. Furthermore we previously found tight correlation of RNA-seq expression and qPCR estimates - in this very same transcriptome[58]. Furthermore, we have actually used the same morphs (AC and SB) and samples in a comparison of the developmental miRNA transcriptome – which reveal that expression of several miRNAs correlates with morph differences[62].

Higher expression of *lyzozyme* and *natterin-like* in SB-charr

The genetic separation in two immunity genes among sympatric morphs in Lake Thingvallavatn[52] prompted us to examine further the expression of *Lyz* and *nattl* that were differentially expressed between morphs.

Both genes are expected to be involved in immune defenses and had higher expression in SB. The substrate of lysozyme[86] is the bacterial cell wall peptidogly-can and it acts directly on Gram-positive bacteria[87]. Lysozyme also promotes the degradation of the outer membrane and therefore indirectly acts also on Gram-negative bacteria[88]. Another gene that caught our attention was *natterin-like*. Natterins were first discovered from the venom gland of the tropical toxic fish species *Thalassophryne nattereri*[75,76], and are found

by sequence similarity in e.g. zebrafish, Atlantic salmon and here in Arctic charr. The predicted Natterin proteins contain a mannose-binding lectin-like domain (Jacalin-domain) and a pore-forming toxin-like domain and can cause edema and pain due to kininogenase activity[75]. Mannose-binding lectins are pathogen recognition proteins (antibodies) and therefore are important for the acute phase response of fish[89,90]. Our data suggest an immune related function of *nattl* genes in charr, as the highest expression was found in skin and kidney. This needs to be verified. It is possible that higher expression of those two genes in SB-charr reflect preparation of juveniles for bottom dwelling habitats, which may be rich in bacteria and challenging for immune systems.

In this study we collapsed contigs into gene or paralog groups for the transcriptome analyses. The disadvantage of this approach is that differential expression in one paralog, can be masked by other related genes that do not differ between groups or have contrasting expression patterns. We looked at this by studying the expression of three paralogs of the *natterin like* genes in different morphs during Arctic charr development, and among tissues of adult AC-charr. The data show that the three *nattl* genes are expressed differentially between the morphs, thus it is not divergence in the expression of one paralog that explains the general *nattl* expression disparity in the transcriptome. Certainly, other scenarios could apply to other genes in the transcriptome.

Expression divergence in craniofacial genes in benthic morphs

A study of the skulls of charr post-hatching embryos and juveniles from Lake Thingvallavatn, showed that some elements of the developing head ossified earlier in SB-charr than in PL-charr[91]. Our new data also demonstrate differences in craniofacial elements between AC- and SB-charr, along a limnetic vs. benthic axis[79]. Based on those differences between benthic and limnetic charr, we investigated further genes with roles in craniofacial development that were differentially expressed in the transcriptome. Guided by this transcriptome we had already found two extra-cellular matrix (ECM) remodeling genes, *Mmp2* and *Sparc* and a conserved co-expression module of genes with known roles in craniofacial morphogenesis, to have higher expression in developing heads of benthic Arctic charr morphs than in limnetic morphs[58,59]. Bioinformatic and qPCR analyses suggest the co-expression module may potentially be affected by quantity of the transcription factor *ETS2*. These studies and the current data confirm the utility of the contrasting developmental transcriptomes for identifying candidate genes with differential expression during head development, as 7 out of 8 candidates were confirmed by qPCR. These genes had consistently higher expression in the developing head of two benthic morphs (SB and LB), and lower in more limnetic fish (AC and PL). This is striking, as three of the morphs (SB, LB and PL) studied are closely related and live in sympatry in Lake Thingvallavatn[52].

We focused on a few targets of Tgf-β and Ahr signaling pathways because of their role in craniofacial morphogenesis and transcriptional connection[92–94]. *Adseverin (Scin)* was one of the top

differentially expressed genes (Table 1) and has roles in rearrangements of the actin cytoskeleton, chondrocyte differentiation and skeletal formation[95,96]. Also, in the transcriptome *Lsr*, *Cldn4* and *Tgfbr2* had higher expression in SB-charr, and we show that higher expression of those genes associated with the benthic morphotype. *Lsr* is a molecular component of tri-cellular tight junctions[97] and has been shown to be suppressed upon Tgf-β1 stimulation[98] in a human cell line. Similarly, *Cldn4*, a tight junction protein with unknown role during embryonic morphogenesis, is a target of the Tgf-β and Ahr signaling pathways[99,100]. Finally, the expression of *Tgfbr2*, encoding a receptor of Tgf-β was slightly but significantly higher in the head of benthic morphs. Previous studies suggest a crucial role of *Tgfbr2* in craniofacial morphogenesis[101].

We also confirmed differential expression of other genes, including two with higher expression in SB-charr. *Mvp* is the predominant component of cytoplasmic ribonucleoprotein structures called vaults[102], which is highly conserved across eukaryotes. The vaults have been something of an enigma, but are implicated in several processes from signal transmission and immune response[103]. The *Jup* gene also showed higher expression in SB-charr. Finally, higher expression of *Vdra*, encoding the vitamin D receptor A, was found in the heads of benthic forms. The receptor regulates mineral homeostasis, osteoblast differentiation and bone metabolism[104].

To summarize, the results show that RNA-sequencing of Aquaculture charr with limnetic craniofacial morphology and small benthic charr can be used to reveal differential expression of genes that associate with limnetic and benthic divergence in craniofacial elements in sympatric charr morphs. It would be interesting if expression of these genes associates with benthic morphology in independently evolved charr populations, or even in other species with similar trophic diversity.

mtDNA function divergence in Arctic charr?

By comparing AC and SB-charr, that represents a small benthic resource morph that has evolved repeatedly in Icelandic stream and pond habitats[51], we hoped to implicate genes and pathways involved in adaptation to these special habitats. The data point to differences between SB and AC-charr in systems related to energy metabolism, as may be expected considering their contrasting life histories. First, there is 2X higher expression of respiratory electron transport chain components in AC compared to SB-charr and 100% more mitochondrial derived reads are found in the AC-charr samples. Note that the direction of divergence is unknown, i.e. whether expression was up in AC or down in SB. Second, many derived candidate-SNPs in genes related to mitochondrial function were at high frequency on the AC branch. For instance in *S100A1*, which has been implicated in mitochondrial regulation in cardiac tissue in humans[105], but its expression is probably not exclusive to this tissue. Third, while the mitochondrial ribosomal genes generally evolve slowly, we do see derived variants at high frequency in the SB and large benthic charr in Lake Thingvallavatn. Specifically, m3411C>T in SB affects a position that is highly conserved among fish, and could affect function of the 16s rRNA. Earlier studies of mitochondrial markers in *S. alpinus* did not find large signals of divergence within Iceland[47,49,52],

probably because they studied other genes. In summary, the results suggest divergence in mitochondrial function, due to the domestication of aquaculture charr and/or possibly reflecting adaptation of the small benthic charr in Lake Thingvallavatn.

The mitochondrion is more than a powerhouse, it integrates metabolism, cell cycle and apoptosis[106]. The number of mitochondria and its functions are known to correlate with environmental attributes. For instance in Antarctic fishes under extreme cold, higher numbers of mitochondria are found in muscle and heart cells[107]. Our data suggest an expression difference between morphs that could reflect differences in total number of mitochondrion, the number of mtDNA copies per mitochondrion or cell, or difference in RNA expression from the mtDNA, possibly due to evolution of mtDNA related to diet and/or temperature[108]. Further work is needed to map out the expression differences of mitochondrial related genes in more SB and anadromous charr morphs (representing the ancestral state). The mtDNA signals could also be investigated in populations along ecological clines (e.g. temperature) or with respect to life history[109].

Conclusions

The data presented here set the stage for future investigations of the molecular and genetic systems involved in the development and divergence of the highly polymorphic and rapidly evolving Arctic charr. The results suggest genetic and expression changes in multiple systems relate to divergence among populations. The data reveal differential expression of two immunological genes between morphs and of several craniofacial developmental genes, that may help sculpture benthic vs. limnetic heads. The genetic data suggest among other things differentiation in the charr mtDNA between morphs. Our broad interest is in how natural selection tweaks genetic regulatory systems, for instance via genetic changes in regulatory sequences or post transcriptional modifiers relating to adaptations. Genetic changes affecting gene expression can be raw material for adaptation, but could also rise in frequency due to reverberations in regulatory cascades[81]. Our specific aim was to cast light on the developmental and population genetics of the unique small benthic charr, typically found in cold springs and small pond habitats in Iceland, particularly those with lava substratum[36,51]. The availability of charr populations at different stages of divergence sets the stage for future genomic studies of the roles of genes, environment and plasticity for shaping this polymorphic species.

Data availability

F1000Research: Dataset 1. Parameters and multiple testing corrected p-values for expression analysis, 10.5256/f1000research.6402.d48005[110]

F1000Research: Dataset 2. qPCR data for tests of expression in charr developing embryos and adult tissues., 10.5256/f1000research.6402.d48006[111]

F1000Research: Dataset 3. qPCR data for tests of expression in charr developing embryo heads., 10.5256/f1000research.6402.d48007[112]

Supporting Information

A **Nattl**

B **Nattl Paralog 1**

C **Nattl Paralog 2**
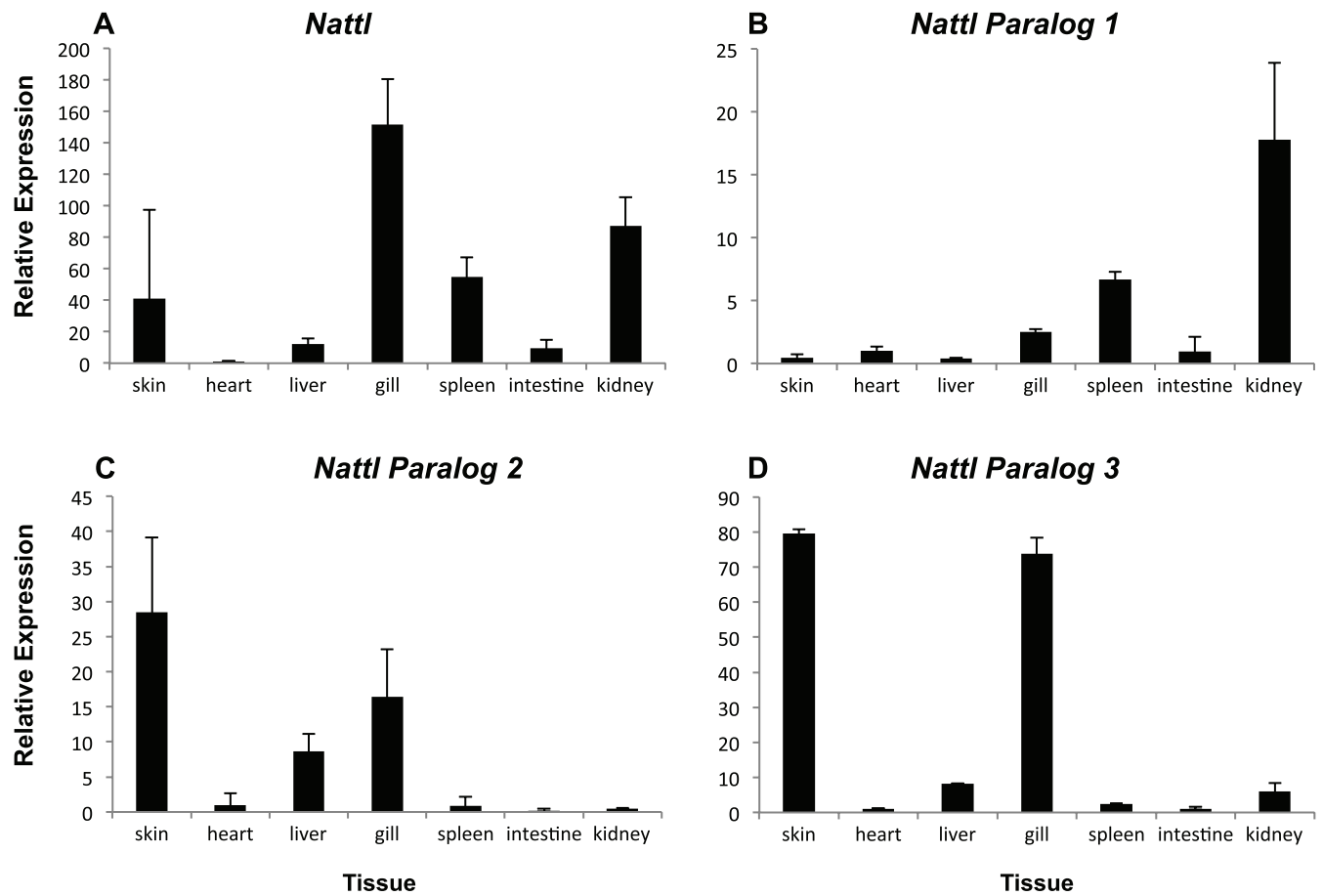
D **Nattl Paralog 3**

S1 Figure. **Relative expression of *nattl* and *nattl 1–3* in tissues of adult AC-charr.** Relative expression of *Natterin* (**A**) & *Natterin paralogs 1–3* (**B**–**D**) within different tissues (skin, heart, liver, gill, spleen, intestine & kidney) of adult aquaculture charr (RT-qPCR); expression plotted for different tissues, relative to heart tissue (lowest expression levels).
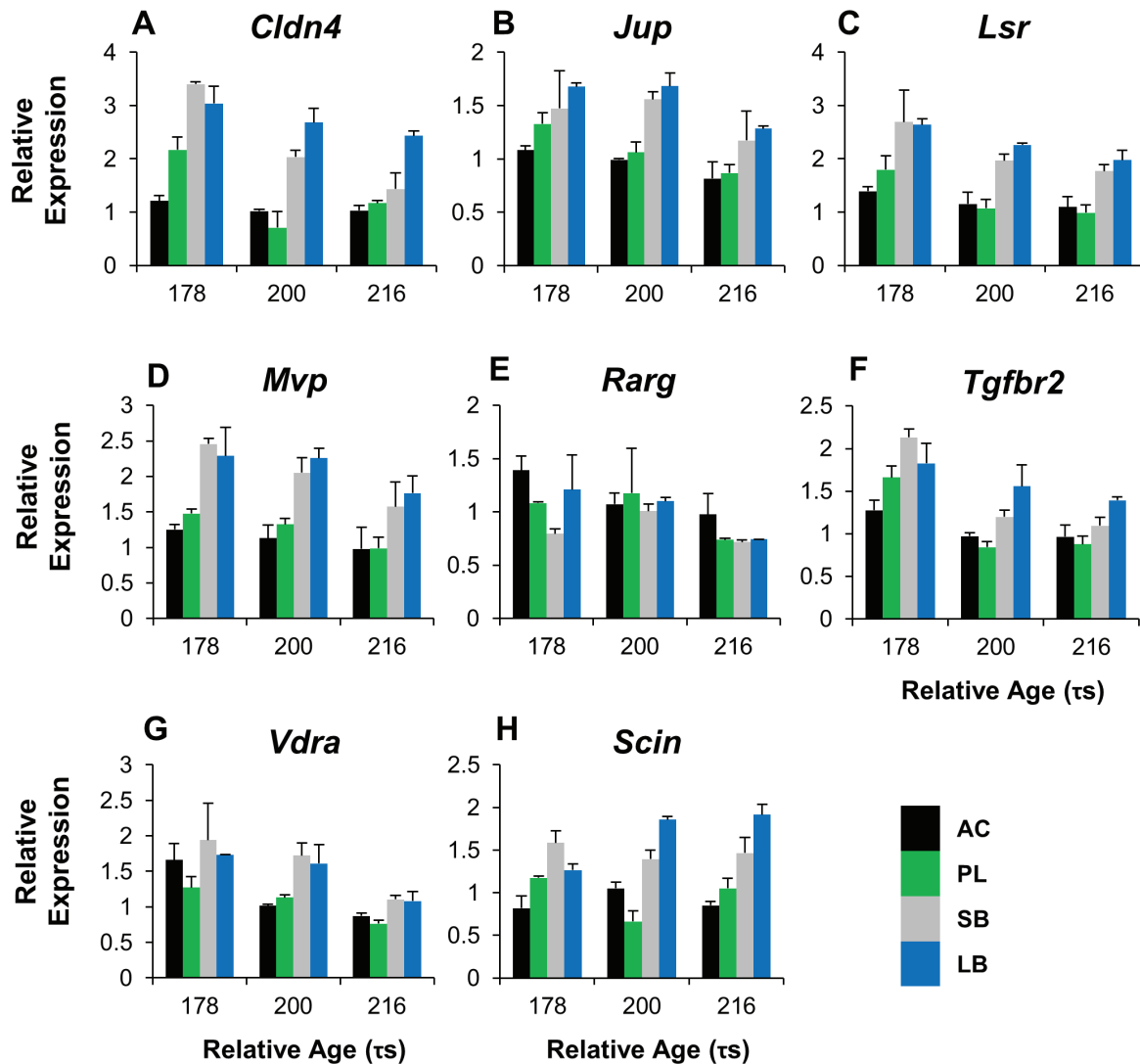
S2 Figure. Relative expression of selected craniofacial candidate genes. Relative expression of 12 candidate genes with characterized craniofacial expression during zebrafish development (ZFIN website) in the head of SB, LB, PL and AC at three time points in development. In the transcriptome data all of the genes had shown higher expression in SB at 200 τs. The expression is normalized to the geometric means of two craniofacial reference genes (*ACTB* and *IF5A1*). Expression is relative to a replicate of AC morph at 200 (τs), set to one. Error bars represent standard deviation calculated from two biological replicates and each biological replicate contains homogenate of six heads.

Supplemental Table S1 A. qPCR primers used in this study.

Gene	Description	Primer Sequence (5'- 3')	Product Size (bp)	PCR Efficiency	Melting Temperature (°C)	Exon Boundary
Actb	Beta Cytoskeletal Actin	F-GAAGATCAAGATCATCGCCC R-CAGACTCGTCGTACTCCTGCT	122	1.95	80.5 ± 0.7	Yes
Alp	Alkaline phosphatase	F-ACAGCATACCTCTGTGGGG R-GGTGGCATGGTTCACACG	177	1.90	85.12 ± 0.5	Yes
Cldn4	Claudin-4	F-GTGCTGTGC CATCCCAAG R-CACCACACAGGTCATCCACA	100	1.98	80.4 ± 0.6	Yes
Cgat2	Chondroitin beta-1,4-N-acetylgalactosaminyltransferase 2	F-GAGAGCCACTTTACTGAGGGG R-GAATGGACGGAAAAGAGTAACG	120	1.98	81.86 ± 0.3	Yes
Cox6b1	Cytochrome c oxidase subunit VIb isoform 1	F-GAGGGTCTACAAATCACTGTGC R-CCTGGAGTCCTACTCATACAAACAT	147	1.93	82.22 ± 0.7	Yes
Ef1α	Eukaryotic Translation Elongation Factor 1 Alpha	F-GAAGATCGGCTATAACCCTGC R-ACCTTCCATCCCTTGAACC	111	1.94	81.36 ± 0.4	Yes
If5a1	Eukaryotic Translation Initiation Factor 5A	F-GGCTTCGTGGTGCTGAAG R-CCATGTGGACCTTAGCGTG	91	1.91	80.76 ± 0.6	Yes
Jup	Junction plakoglobin	F-CACAGCAGACATACCAGGATG G R-CTGGCGATCTCTCCCCTGTT	109	1.97	81.0 ± 0.3	Yes
Krtap4–3	Keratin-associated protein 4–3	F-GCGGGACATCTACACTGCTTA R-AGAAGGCTAAAGTCTTAGTGACTATC	151	1.89	81.88 ± 0.6	Yes
Lsr	Lipolysis-stimulated lipoprotein receptor	F-TGCTGTCACTCTGGGCGA R-CCGTCTGGGCAAGGTTCA G	80	1.91	80.77 ± 0.5	Yes
Lyz	Lysozyme	F-TTCCAGATCAACAGCCGCTA R-GATCGCCACTGTGATGTCAT	111	1.94	81.87 ± 0.7	Yes
Mvp	Major vault protein	F-ACCAACTCCCAGGAGGCT R-CCTCTCCAGACGACCACG	75	1.97	78.93 ± 0.3	Yes
Nattl	Natterin-like protein	F-GTGAAAGTCACCTGCATGAATG R-CATCTCTCCTTTGTGGATACCC	104	1.98	78.81 ± 0.8	No
Nattl-1	Natterin-like protein paralog-1	F-AATCCGTGTCCTACCACAATGA R-GGTGTGTCGGTCAAAGCA	135	1.77	78.03 ± 0.1	No
Nattl-2	Natterin-like protein paralog-1	F-TGAAATVTVTGTCTCATCACAAC R-GGATCTGGTCGAGGTGGC	163	1.72	80.50 ± 0.2	No
Nattl-3	Natterin-like protein paralog-1	F-GTGACATCCGTTTCTCACCAG R-GATGTGTCGGTCAAAGCG	138	1.77	79.12 ± 0.2	No
Ndub6	NADH dehydrogenase 1 beta subcomplex subunit 6	F-TGGTGGAGTGTTCGCCTT R-CTCTCTGGGAGGTCTGGAA	171	1.89	82.40 ± 0.3	Yes
Parp6	Poly (ADP-Ribose) Polymerase Family, Member 6	F-CCGTATGAATACCGTTCCACAGG R-CACCCAGATGTTGCCGTGCTT	147	1.93	81.87 ± 0.7	Yes

Supplemental Table S1 B.

Gene	Description	Primer Sequence (5'- 3')	Product Size (bp)	PCR Efficiency	Melting Temperature (°C)	Exon Boundary
Rarg	Retinoic acid receptor gamma-A	F-AAGGCGAGCCCCTTCTTC R-TGCTCTGGGTCTCCACCG	82	1.92	78.62 ± 0.3	Yes
Scin	Scinderin/Adseverin	F-CACCTGATCCCAGACATCCAA R-CCTCACTCAACAACCTCGC	136	1.90	83.24 ± 0.7	No
Tgfbr2	TGF-beta receptor type-2	F-CTGCTCCGAGGACGAGTG R-ACCGACACCACCTGGGAG	72	1.93	79.02 ± 0.5	Yes
Ubl5	Ubiquitin-like protein 5	F-AATAAGGATGATTGAGGTGGTTTG R-ATGAGCTTCTTCAGGTCTCC	99	1.95	78.44 ± 0.3	Yes
Ub2l3	Ubiquitin-Conjugating Enzyme E2L 3	F-CGAGAAGGGACAGGTGTGTC R-ACCAACGCAATCAGGGACT	96	1.93	79.62 ± 0.3	Yes
Vdra	Vitamin D3 receptor A	F-CGTCACCAAGGCGGGTCA R-TGGAGCTTG AGTTTCTTCAGGC	81	1.93	78.12 ± 0.3	Yes

Supplemental Table S2 A. Verification of candidate polymorphisms. Primer sequences, melting temperatures and primary data.

Sequence	Position	Forward primer	Reverse primer	Tm forward	Tm reverse	Paralogs
NC_000861.1	1829	GTGCCTCAGACCCACCTAGA	TCTGTCGCCCGTACTAAGGT	60.26	59.76	No
NC_000861.1	3119	GGCCAGAGTAAACACCGAGA	CCTGGATTACTCCGGTCTGA	60.25	60.07	No
NC_000861.1	3411	GGCCAGAGTAAACACCGAGA	CCTGGATTACTCCGGTCTGA	60.25	60.07	No
NC_000861.1	8876	GACGTCCTTCACTCCTGAGC	GGGCTCATAAACTGGTCGAA	59.99	60.07	No
NC_000861.1	15240	ACCCTAAAACCGAACGATCC	TGGCTAGGAAGAGTCCGGTA	60.19	59.83	No
SS2U034121	233	CTCAACGTGCTTGACCAGTG	CCCTTACCCTCCAGGATCTC	60.5	59.89	Yes
SS2U054644	1037	AAGGACGGCCACTATGGTCT	GGGGCATAGAGTGCACAGG	60.9	61.65	Yes
SS2U054644	1188	TCAGAGATAGTGAAGAAGATGCTG	CGTACTTGATAAGACCTGTCGGTA	57.92	59.62	No
SS2U054644	1283	TCAGAGATAGTGAAGAAGATGCTG	CGTACTTGATAAGACCTGTCGGTA	57.92	59.62	No
SS2U055283	1822	TGTGTGAGGTGGTTGAGGAG	GGGTCATTGCTCCCTACAGA	59.7	60.07	No
SS2U055923	615	GTGGACCCAGAGGATGAGAA	AGAACCTGCTCCCAGTTTGA	60.05	59.84	No
SS2U058906	350	GCCAAAACCTCCACAATGAT	AACTGGCCTTCCAGATCAGA	59.8	59.8	Yes/No

Paralogs: indicates whether the PCR and sequencing yielded mixed products, indicative of paralogous genes.

Supplemental Table S2 B.

Sequence	Genome contig	Gene name	Position	Ref	Var	Freq_AC	Freq_SB	FreqP_PL	FreqP_SB	FreqP_LB
NC_000861.1	n.a.	12S ribosomal RNA	1829	G	A	0 / 53	77 / 81	0 / 6	3 / 4	1 / 8
NC_000861.1	n.a.	16S ribosomal RNA	3119	A	T	46 / 87	18 / 28	0 / 8	0 / 8	0 / 8
NC_000861.1	n.a.	16S ribosomal RNA	3411	C	T	0 / 119	26 / 33	0 / 8	5 / 8	1 / 8
NC_000861.1	n.a.	tRNA-Lys	8876	C	A	73 / 779	74 / 352	0 / 4	0 / 4	n.a.
NC_000861.1	n.a.	NADH dehydrogenase 6	15240	G	A	2 / 3608	137 / 2702	2 / 4	0 / 4	n.a.
SS2U034121	AGKD01052493.1	Eukaryotic translation initiation factor 4 gamma 2	233	C	T	0 / 95	22 / 40	2 / 4	2 / 2	n.a.
SS2U054644	AGKD01031893.1	Uroporphyrinogen decarboxylase	1037	G	A	28 / 33	0 / 56	n.a.	n.a.	n.a.
SS2U054644	AGKD01031893.1	Uroporphyrinogen decarboxylase	1188	C	T	0 / 53	19 / 25	4 / 4	4 / 4	n.a.
SS2U054644	AGKD01031893.1	Uroporphyrinogen decarboxylase	1283	G	A	4 / 60	12 / 15	4 / 4	4 / 4	n.a.
SS2U055283	AGKD01013777.1	DNA2-like helicase	1822	G	A	1 / 65	25 / 50	3 / 4	n.a.	n.a.
SS2U055923	AGKD01022586.1	Bystin	615	G	A	106 / 109	7 / 190	0 / 4	n.a.	n.a.
SS2U058906	AGKD01005918.1	Mid1-interacting protein 1-like	350	G	T	0 / 49	67 / 68	4 / 4	4 / 4	n.a.

Sequence: name of the genebank sequence or EST-contig used as reference for mapped reads.

Genome contig: name of salmon genome (ICSASG_v1) contig with best sequence match to the respective EST-contig.

Ref: Reference variant.

Var: The derived variant.

Freq_AC and Freq_SB: Frequency of variant reads as fraction of total numbers of reads mapped in Aquaculture (AC) or Small benthic (SB).

FreqP: The frequency of variant in genotyping by PCR and direct sequencing, as a fraction of total number of chromosomes sequenced.

Supplemental Table S3. Mapping of Illumina reads to *S. salar* EST data. Numbers of reads aligning to salmon reference for each sample.

Alignment per read	SB 141	SB 163	SB 200	SB 433	AC 141	AC 163	AC 200	AC 433
0	33088778	30492314	27175901	25569628	32159386	30051365	31267710	28563169
1	6979368	11791558	11449549	11058555	11599602	11320997	11027195	10650748
2	2742358	4021683	3814418	3734404	4328402	4523686	3959198	3655786
3	2099068	2964994	2748108	2651522	3111277	3332577	2878729	2515303
4	1228292	1777846	1720902	1968251	1977738	2182392	1929818	1980420
5	914704	1317556	1284262	1434314	1471739	1679277	1447604	1426744
6	645264	946579	938290	1087959	1001350	1083025	1045157	1081063
7	425856	595785	578175	726290	657220	750523	690286	735351
8	293065	428003	424426	590100	530040	591332	527821	579860
9	206205	319401	334861	455838	296169	334264	387901	485653
10+	749074	1419362	1761275	3041930	1092980	1189781	1967857	3294222
Total reads	49372032	56075081	52230167	52318791	58225903	57039219	57129276	54968319

Supplemental Table S4. ANOVAs on qPCR data. Expression of nine genes was analyzed in whole SB and AC-charr embryos, at two developmental timepoints (161 and 200 day degrees).

Gene	Term	Df	F value	p value	Significance	FDR RNA-seq
Alp	Morph	1	13.4797	0.0214	*	0.0697
	Time	1	14.9526	0.0180	*	0.0012
	M x T	1	3.9519	0.1177	.	
Cgat2	Morph	1	0.0257	0.8804	.	0.0035
	Time	1	1.5141	0.2859	.	0.3312
	M x T	1	0.1866	0.6880	.	
Cox6B1	Morph	1	0.0898	0.7793	.	0.0580
	Time	1	3.8312	0.1219	.	0.6320
	M x T	1	0.7359	0.4393	.	
Krtap4–3	Morph	1	30.0255	0.0054	**	0.0121
	Time	1	0.3902	0.5661	.	0.2784
	M x T	1	4.5225	0.1006	.	
Lyz	Morph	1	64.1566	0.0013	**	0.0406
	Time	1	1.0390	0.3657	.	0.0005
	M x T	1	1.2026	0.3344	.	
Nattl	Morph	1	8.1148	0.0465	*	7.718e-07
	Time	1	14.6659	0.0186	*	6.714e-14
	M x T	1	0.2958	0.6154	.	
Ndub6	Morph	1	0.7447	0.4368	.	0.0982
	Time	1	7.3316	0.0537	.	0.6698
	M x T	1	0.2269	0.6587	.	
Parp6	Morph	1	11.2682	0.0284	*	0.1076
	Time	1	0.7393	0.4384	.	0.3789
	M x T	1	0.2343	0.6537	.	
Ubl5	Morph	1	1.1420	0.3454	.	0.0587
	Time	1	0.2434	0.6476	.	0.0025
	M x T	1	0.3974	0.5627	.	

Significance: $p > 0.05$; * $p < 0.05$; ** $p < 0.01$.

FDR RNA-seq: indicates significance of Morph and Time effects in the transcriptome data.

Supplemental Table S5. ANOVA for relative expression levels of *Natterin-like* and *Natterin-like* paralogs 1–3 in Arctic charr whole embryos (among SB, AC and PL morphs) and tissues from adult AC-charr.

Gene	Term	Df	F value	p value	Significance
Nattl	Morph	2	11.5515	0.0002	***
	Time	5	8.3202	3.99e-05	***
	M x T	9	4.4758	0.0007	***
Nattl1	Morph	2	19.4070	0.0001	***
	Time	3	5.9346	0.0089	**
	M x T	5	4.5761	0.0126	*
Nattl2	Morph	2	14.2921	0.0005	***
	Time	3	15.0463	0.0001	***
	M x T	5	3.2462	0.0404	*
Nattl3	Morph	2	34.4888	6.33e-06	***
	Time	3	4.4204	0.0238	*
	M x T	5	4.1843	0.0174	*
Nattl	Tissue	6	15.468	1.42e-08	***
Nattl1	Tissue	6	12.022	0.0002	***
Nattl2	Tissue	6	7.6811	0.0011	**
Nattl3	Tissue	6	46.182	8.89e-06	***

Significance: $p > 0.05$; * $p < 0.05$; ** $p < 0.01$.

Supplemental Table S6. Gene Ontology analyses of derived SNPs in SB-charr.

Category	Observed	In category	TERM	FDR adjusted p-value
GO:0006412	24	189	translation	4.34E-006
GO:0006396	8	32	RNA processing	0.0016
GO:0006414	6	19	translational elongation	0.0038
GO:0006313	5	20	transposition, DNA-mediated	0.0498
GO:0015074	5	21	DNA integration	0.0510
GO:0006260	6	35	DNA replication	0.0679
GO:0055114	20	285	oxidation-reduction process	0.0679

Supplemental Table S7. Predicted effect of SNP-candidates differing in frequency between charr morphs.

Effect on transcribed region	Uni_SB	Uni_AC	Rep_SB	Rep_AC
5´prime	32	19	35	24
Synonymous	232	179	176	113
Non-synonymous	112	72	81	72
3´prime	147	123	59	74

From RNA-reads that mapped to one (Uni) or more (Rep) *S. salar* ESTs.

The candidate SNPs frequencies differ more than 50% between SB and AC-charr, summarized by which morph with higher frequency of the derived allele.

References

1. Gould SJ: **Ontogeny and Phylogeny**. Harvard University Press, 1977.
   Reference Source

2. Jacob F: **Evolution and tinkering.** *Science.* 1977; **196**(4295): 1161–1166.
   PubMed Abstract | Publisher Full Text

3. Stern DL, Orgogozo V: **The loci of evolution: how predictable is genetic evolution?** *Evolution.* 2008; **62**(9): 2155–77.
   PubMed Abstract | Publisher Full Text | Free Full Text

4. Stern DL, Orgogozo V: **Is genetic evolution predictable?** *Science.* 2009; **323**(5915): 746–751.
   PubMed Abstract | Publisher Full Text | Free Full Text

5. Rockman MV: **The QTN program and the alleles that matter for evolution: all that's gold does not glitter.** *Evolution.* 2012; **66**(1): 1–17.
   PubMed Abstract | Publisher Full Text | Free Full Text

6. Abzhanov A, Protas M, Grant BR, *et al.*: *Bmp4* **and morphological variation of beaks in Darwin's finches.** *Science.* 2004; **305**(5689): 1462–1465.
   PubMed Abstract | Publisher Full Text

7. Albertson RC, Kocher TD: **Genetic and developmental basis of cichlid trophic diversity.** *Heredity (Edinb).* 2006; **97**(3): 211–221.
   PubMed Abstract | Publisher Full Text

8. Cresko WA, McGuigan KL, Phillips PC, *et al.*: **Studies of threespine stickleback developmental evolution: progress and promise.** *Genetica.* 2007; **129**(1): 105–126.
   PubMed Abstract | Publisher Full Text

9. Giger T, Excoffier L, Day PJ, *et al.*: **Life history shapes gene expression in salmonids.** *Curr Biol.* 2006; **16**(8): R281–2.
   PubMed Abstract | Publisher Full Text

10. Gross JB, Furterer A, Carlson BM, *et al.*: **An integrated transcriptome-wide analysis of cave and surface dwelling** *Astyanax mexicanus*. *PLoS One.* 2013; **8**(2): e55659.
    PubMed Abstract | Publisher Full Text | Free Full Text

11. Lanes CF, Bizuayehu TT, de Oliveira Fernandes JM, *et al.*: **Transcriptome of Atlantic Cod (***Gadus morhua L.***) early embryos from farmed and wild broodstocks.** *Mar Biotechnol (NY).* 2013; **15**(6): 677–694.
    PubMed Abstract | Publisher Full Text

12. Bougas B, Audet C, Bernatchez L: **The influence of parental effects on transcriptomic landscape during early development in brook charr (***Salvelinus fontinalis***, Mitchill).** *Heredity (Edinb).* 2013; **110**(5): 484–491.
    PubMed Abstract | Publisher Full Text | Free Full Text

13. Micallef G, Bickerdike R, Reiff C, *et al.*: **Exploring the transcriptome of Atlantic salmon (***Salmo salar***) skin, a major defense organ.** *Mar Biotechnol (NY).* 2012; **14**(5): 559–569.
    PubMed Abstract | Publisher Full Text

14. Wang S, Furmanek T, Kryvi H, *et al.*: **Transcriptome sequencing of Atlantic salmon (***Salmo salar L.***) notochord prior to development of the vertebrae provides clues to regulation of positional fate, chordoblast lineage and mineralisation.** *BMC Genomics.* 2014; **15**: 141.
    PubMed Abstract | Publisher Full Text | Free Full Text

15. Qian X, Ba Y, Zhuang Q, *et al.*: **RNA-Seq technology and its application in fish transcriptomics.** *OMICS.* 2014; **18**(2): 98–110.
    PubMed Abstract | Publisher Full Text | Free Full Text

16. Norman JD, Ferguson MM, Danzmann RG: **Transcriptomics of salinity tolerance capacity in Arctic charr (***Salvelinus alpinus***): a comparison of gene expression profiles between divergent QTL genotypes.** *Physiol Genomics.* 2014; **46**(4): 123–37.
    PubMed Abstract | Publisher Full Text | Free Full Text

17. Derome N, Duchesne P, Bernatchez L: **Parallelism in gene transcription among sympatric lake whitefish (***Coregonus clupeaformis* Mitchill) ecotypes.** *Mol Ecol.* 2006; **15**(5): 1239–49.
    PubMed Abstract | Publisher Full Text

18. Filteau M, Pavey SA, St-Cyr J, *et al.*: **Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish.** *Mol Biol Evol.* 2013; **30**(6): 1384–96.
    PubMed Abstract | Publisher Full Text

19. Domazet-Lošo T, Tautz D: **A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns.** *Nature.* 2010; **468**(7325): 815–8.
    PubMed Abstract | Publisher Full Text

20. Bozinovic G, Sit TL, Hinton DE, *et al.*: **Gene expression throughout a vertebrate's embryogenesis.** *BMC genomics.* 2011; **12**: 132.
    PubMed Abstract | Publisher Full Text | Free Full Text

21. Skúlason S, Smith TB: **Resource polymorphisms in vertebrates.** *Trends Ecol Evol.* 1995; **10**(9): 366–370.
    PubMed Abstract | Publisher Full Text

22. Snorrason SS, Skúlason S: **Adaptive Speciation in Northern Freshwater Fishes**. In Ulf Dieckmann, Michael Doebeli, Johan A. J. Metz, and Diethard Tautz, editors, *Adaptive speciation.* Cambridge University Press, Cambridge, chapter 10, 2004; 210–228.
    Publisher Full Text

23. Klemetsen A: **The Charr Problem Revisited: Exceptional Phenotypic Plasticity Promotes Ecological Speciation in Postglacial Lakes.** *Freshwater Rev.* 2010;
    3(1): 49–74.
    Publisher Full Text

24. Bernatchez L, Renaut S, Whiteley AR, *et al.*: **On the origin of species: insights from the ecological genomics of lake whitefish.** *Philos Trans R Soc Lond B Biol Sci.* 2010; **365**(1547): 1783–800.
    PubMed Abstract | Publisher Full Text | Free Full Text

25. Merilä J: **Nine-spined stickleback (***Pungitius pungitius***): an emerging model for evolutionary biology research.** *Ann N Y Acad Sci.* 2013; **1289**: 18–35.
    PubMed Abstract | Publisher Full Text

26. Fraser DJ, Weir LK, Bernatchez L, *et al.*: **Extent and scale of local adaptation in salmonid fishes: review and meta-analysis.** *Heredity (Edinb).* 2011; **106**(3): 404–20.
    PubMed Abstract | Publisher Full Text | Free Full Text

27. Allendorf FW, Thorgaard GH: **Tetraploidy and the Evolution of Salmonid Fishes**. In Bruce J. Turner, editor, *Evolutionary Genetics of Fishes.* Springer US, 1984; 1–53.
    Publisher Full Text

28. Crête-Lafrenière A, Weir LK, Bernatchez L: **Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling.** *PLoS One.* 2012; **7**(10): e46662.
    PubMed Abstract | Publisher Full Text | Free Full Text

29. Davidson WS, Koop BF, Jones SJ, *et al.*: **Sequencing the genome of the Atlantic salmon (***Salmo salar***).** *Genome Biol.* 2010; **11**(9): 403.
    PubMed Abstract | Free Full Text

30. Moghadam HK, Ferguson MM, Danzmann RG: **Whole genome duplication: Challenges and considerations associated with sequence orthology assignment in Salmoninae.** *J Fish Biol.* 2011; **79**(3): 561–574.
    PubMed Abstract | Publisher Full Text

31. Macqueen DJ, Johnston IA: **A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification.** *Proc Biol Sci.* 2014; **281**(1778): 20132881.
    PubMed Abstract | Publisher Full Text | Free Full Text

32. Berthelot C, Brunet F, Chalopin D, *et al.*: **The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates.** *Nat Commun.* 2014; **5**: 3657.
    PubMed Abstract | Publisher Full Text | Free Full Text

33. Vijay N, Poelstra JW, Künstner A, *et al.*: **Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive** *in silico* **assessment of RNA-seq experiments.** *Mol Ecol.* 2013; **22**(3): 620–34.
    PubMed Abstract | Publisher Full Text

34. Noakes DLG: **Charr truth: Sympatric differentiation in** *Salvelinus* **species.** *Environ Biol Fishes.* 2008; **83**(1): 7–15.
    Publisher Full Text

35. Adams CE, Fraser D, Wilson AJ, *et al.*: **Patterns of phenotypic and genetic variability show hidden diversity in Scottish Arctic charr.** *Ecol Freshwater Fish.* 2007; **16**(1): 78–86.
    Publisher Full Text

36. Kristjánsson BK, Malmquist HJ, Ingimarsson F, *et al.*: **Relationships between lake ecology and morphological characters in Icelandic Arctic charr,** *Salvelinus alpinus*. *Biol J Linn Soc.* 2011; **103**(4): 761–771.
    Publisher Full Text

37. Woods PJ, Skúlason S, Snorrason SS, *et al.*: **Intraspecific diversity in Arctic charr,** *Salvelinus alpinus*, **in Iceland: II. Which environmental factors influence resource polymorphism in lakes?** *Evolutionary Ecol Res.* 2012; **14**(8): 993–1013.
    Reference Source

38. Kristjánsson BK, Skúlason S, Snorrason SS, *et al.*: **Fine-scale parallel patterns in diversity of small benthic Arctic charr (***Salvelinus alpinus***) in relation to the ecology of lava/groundwater habitats.** *Ecol Evol.* 2012; **2**(6): 1099–112.
    PubMed Abstract | Publisher Full Text | Free Full Text

39. Skúlason S, Noakes DL, Snorrason SS: **Ontogeny of trophic morphology in four sympatric morphs of Arctic charr** *Salvelinus alpinus* **in Thingvallavatn, Iceland.** *Biol J Linn Soc.* 1989; **38**(3): 281–301.
    Publisher Full Text

40. Skúlason S, Snorrason SS, Ota D, *et al.*: **Genetically based differences in foraging behaviour among sympatric morphs of Arctic charr (Pisces: Salmonidae).** *Animal Behaviour.* 1993; **45**(6): 1179–1192.
    Publisher Full Text

41. Skúlason S, Snorrason SS, Noakes DLG, *et al.*: **Genetic basis of life history variations among sympatric morphs of Arctic char** *Salvelinus alpinus*. *Can J Fish Aquat Sci.* 1996; **53**(8): 1807–1813.
    Publisher Full Text

42. Parsons KJ, Skúlason S, Ferguson M: **Morphological variation over ontogeny and environments in resource polymorphic Arctic charr (***Salvelinus alpinus***).** *Evol Dev.* 2010; **12**(3): 246–257.
    PubMed Abstract | Publisher Full Text

43. Parsons KJ, Sheets HD, Skúlason S, *et al.*: **Phenotypic plasticity, heterochrony and ontogenetic repatterning during juvenile development of divergent Arctic charr (***Salvelinus alpinus***).** *J Evolutionary Biol.* 2011; **24**(8): 1640–1652.
    PubMed Abstract | Publisher Full Text

44. Sandlund TO, Gunnarsson K, Jónasson PM, *et al.*: **The Arctic charr *Salvelinus alpinus* in Thingvallavatn.** *Oikos.* 1992; **64**(1/2): 305–351.
Publisher Full Text

45. Snorrason SS, Skúlason S, Jonsson B, *et al.*: **Trophic specialization in Arctic charr *Salvelinus alpinus* (Pisces; Salmonidae): morphological divergence and ontogenetic niche shifts.** *Biol J Linn Soc.* 1994; **52**(1): 1–18.
Publisher Full Text

46. Magnusson KP, Ferguson MM: **Genetic analysis of four sympatric morphs of Arctic charr, *Salvelinus alpinus*, from Thingvallavatn, Iceland.** *Environ Biol Fishes.* 1987; **20**(1): 67–73.
Publisher Full Text

47. Danzmann RG, Ferguson MM, Skulason S, *et al.*: **Mitochondrial DNA diversity among four sympatric morphs of Arctic charr, *Salvelinus alpinus* L., from Thingvallavatn, Iceland.** *J Fish Biol.* 1991; **39**(5): 649–659.
Publisher Full Text

48. Pálsson S, Árnason E: **Sequence variation for cytochrome b genes of three salmonid species from Iceland.** *Aquaculture.* 1994; **128**(1–2): 29–39.
Publisher Full Text

49. Volpe JP, Ferguson MM: **Molecular genetic examination of the polymorphic Arctic charr *Salvelinus alpinus* of Thingvallavatn, Iceland.** *Mol Ecol.* 1996; **5**(6): 763–72.
PubMed Abstract | **Publisher Full Text**

50. Wilson AJ, Gíslason D, Skúlason S, *et al.*: **Population genetic structure of Arctic charr, *Salvelinus alpinus* from northwest Europe on large and small spatial scales.** *Mol Ecol.* 2004; **13**(5): 1129–42.
PubMed Abstract | **Publisher Full Text**

51. Kapralova KH, Morrissey MB, Kristjánsson BK, *et al.*: **Evolution of adaptive diversity and genetic connectivity in Arctic charr (*Salvelinus alpinus*) in Iceland.** *Heredity (Edinb).* 2011; **106**(3): 472–487.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

52. Kapralova KH, Gudbrandsson J, Reynisdottir S, *et al.*: **Differentiation at the *MHCIIα* and *Cath2* loci in sympatric *Salvelinus alpinus* resource morphs in Lake Thingvallavatn.** *PLoS One.* 2013; **8**(7): e69402.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

53. Saemundsson K: **Geology of the Thingvallavatn Area.** *Oikos.* 1992; **64**(1/2): 40–68.
Publisher Full Text

54. Macqueen DJ, Kristjánsson BK, Paxton CG, *et al.*: **The parallel evolution of dwarfism in Arctic charr is accompanied by adaptive divergence in mTOR-pathway gene expression.** *Mol Ecol.* 2011; **20**(15): 3167–84.
PubMed Abstract | **Publisher Full Text**

55. Doiron S, Bernatchez L, Blier PU: **A comparative mitogenomic analysis of the potential adaptive value of Arctic charr mtDNA introgression in brook charr populations (*Salvelinus fontinalis* Mitchill).** *Mol Biol Evol.* 2002; **19**(11): 1902–9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

56. Miller W, Schuster SC, Welch AJ, *et al.*: **Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change.** *Proc Natl Acad Sci U S A.* 2012; **109**(36): E2382–90.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

57. Svavarsson E: **Árangur í kynbótum á bleikju og næstu skref [reference in Icelandic].** In *Fræðaþing landbúnaðarsins 4.* 2007; 121–125.
Reference Source

58. Ahi EP, Gudbrandsson J, Kapralova KH, *et al.*: **Validation of Reference Genes for Expression Studies during Craniofacial Development in Arctic Charr.** *PloS One.* 2013; **8**(6): e66389.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

59. Ahi EP, Kapralova KH, Pálsson A, *et al.*: **Transcriptional dynamics of a conserved gene expression network associated with craniofacial divergence in Arctic charr.** *EvoDevo.* 2014; **5**(1): 40.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

60. Snorrason SS, Skúlason S, Sandlund OT, *et al.*: **Shape polymorphism in Arctic charr, *Salvelinus alpinus*.** *Physiol Ecol Japan.* 1989; **1**: 393–404.
Reference Source

61. Gorodilov YN: **Description of the early ontogeny of the Atlantic salmon, *Salmo salar*, with a novel system of interval (state) identification.** *Environ Biol Fishes.* 1996; **47**(2): 109–127.
Publisher Full Text

62. Kapralova KH, Franzdóttir SR, Jónsson H, *et al.*: **Patterns of MiRNA expression in Arctic Charr development.** *PLoS One.* 2014; **9**(8): e106084.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

63. Di Génova A, Aravena A, Zapata L, *et al.*: **SalmonDB: a bioinformatics resource for *Salmo salar* and *Oncorhynchus mykiss*.** *Database (Oxford).* 2011; **2011**: bar050.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

64. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics.* 2011; **12**(1): 323.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

65. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–40.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

66. R Core Team. **R: A Language and Environment for Statistical Computing.** 2014.

67. Young MD, Wakefield MJ, Smyth GK, *et al.*: **Gene ontology analysis for RNA-seq: accounting for selection bias.** *Genome Biol.* 2010; **11**(2): R14.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

68. Untergasser A, Cutcutache I, Koressaar T, *et al.*: **Primer3–new capabilities and interfaces.** *Nucleic Acids Res.* 2012; **40**(15), e115.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

69. Bustin SA, Benes V, Garson JA, *et al.*: **The MIQE guidelines: *M*inimum *I*nformation for publication of *Q*uantitative real-time PCR *E*xperiments.** *Clin Chem.* 2009; **55**(4): 611–22.
PubMed Abstract | **Publisher Full Text**

70. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method.** *Methods.* 2001; **25**(4): 402–8.
PubMed Abstract | **Publisher Full Text**

71. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

72. Koboldt DC, Zhang Q, Larson DE, *et al.*: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome Res.* 2012; **22**(3): 568–76.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

73. Larkin MA, Blackshields G, Brown NP, *et al.*: **Clustal W and Clustal X version 2.0.** *Bioinformatics.* 2007; **23**(21): 2947–2948.
PubMed Abstract | **Publisher Full Text**

74. Nicholas KB, Nicholas HB, Deerfield DW: **GeneDoc: Analysis and Visualization of Genetic Variation.** *EMBNEW.NEWS.* 1997; **4**(14).
Reference Source

75. Magalhães GS, Lopes-Ferreira M, Junqueira-de Azevedo ILM, *et al.*: **Natterins, a new class of proteins with kininogenase activity characterized from *Thalassophryne nattereri* fish venom.** *Biochimie.* 2005; **87**(8): 687–99.
PubMed Abstract | **Publisher Full Text**

76. Magalhães GS, Junqueira-de-Azevedo IL, Lopes-Ferreira M, *et al.*: **Transcriptome analysis of expressed sequence tags from the venom glands of the fish *Thalassophryne nattereri*.** *Biochimie.* 2006; **88**(6): 693–9.
PubMed Abstract | **Publisher Full Text**

77. Sprague J, Bayraktaroglu L, Clements D, *et al.*: **The Zebrafish Information Network: the zebrafish model organism database.** *Nucleic Acids Res.* 2006; **34**(Database issue): D581–D585.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

78. Eiriksson GM, Skulason S, Snorrason SS: **Heterochrony in skeletal development and body size in progeny of two morphs of Arctic charr from Thingvallavatn, Iceland.** *J Fish Biol.* 1999; **55**(sA): 175–185.
Publisher Full Text

79. Kapralova KH, Jónsson ZO, Pálsson A, *et al.*: **Bones in motion: Ontogeny of craniofacial development in sympatric Arctic charr morphs.** *Developmental Dynamics, under revision.* 2015.

80. Jonsson B, Skúlason S, Snorrason SS, *et al.*: **Life History Variation of Polymorphic Arctic Charr (*Salvelinus alpinus*) in Thingvallavatn, Iceland.** *Can J Fish Aquat Sci.* 1988; **45**(9): 1537–1547.
Publisher Full Text

81. Palsson A, Wesolowska N, Reynisdóttir S, *et al.*: **Naturally occurring deletions of hunchback binding sites in the *even-skipped* stripe 3+7 enhancer.** *PLoS One.* 2014; **9**(5); e91924.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

82. Zou C, Lehti-Shiu MD, Thomashow M, *et al.*: **Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*.** *PLoS Genet.* 2009; **5**(7): e1000581.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

83. Jantzen SG, Sanderson DS, von Schalburg KR, *et al.*: **A 44K microarray dataset of the changing transcriptome in developing Atlantic salmon (*Salmo salar* L.).** *BMC Res Notes.* 2011; **4**: 88.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

84. Drivenes Ø, Taranger GL, Edvardsen RB: **Gene expression profiling of Atlantic cod (*Gadus morhua*) embryogenesis using microarray.** *Mar Biotechnol (NY).* 2012; **14**(2): 167–76.
PubMed Abstract | **Publisher Full Text**

85. Piasecka B, Lichocki P, Moretti S, *et al.*: **The hourglass and the early conservation models--co-existing patterns of developmental constraints in vertebrates.** *PLoS Genet.* 2013; **9**(4): e1003476.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

86. Fleming A: **On a Remarkable Bacteriolytic Element Found in Tissues and Secretions.** *Proceedings of the Royal Society of London. Series B.* 1922; **93**(653): 306–317.
Publisher Full Text

87. Chipman DM, Sharon N: **Mechanism of lysozyme action.** *Science.* 1969; **165**(3892): 454–465.
PubMed Abstract | **Publisher Full Text**

88. Subramanian S, MacKinnon SL, Ross NW: **A comparative study on innate immune parameters in the epidermal mucus of various fish species.** *Comp Biochem Physiol B Biochem Mol Biol.* 2007; **148**(3): 256–63.
PubMed Abstract | **Publisher Full Text**

89. Magnadóttir B: **Innate immunity of fish (overview).** *Fish Shellfish Immunol.* 2006;

20(2): 137–51.
PubMed Abstract | **Publisher Full Text**

90. Magnadottir B: **Immunological control of fish diseases.** *Mar Biotechnol (NY).* 2010; **12**(4): 361–79.
PubMed Abstract | **Publisher Full Text**

91. Eiriksson GM: **Heterochrony in bone development and growth in two morphs of Arctic charr (*Salvelinus alpinus*) from Thingvallavatn, Iceland**. Master's thesis, University of Iceland, 1999.

92. Chai Y, Ito Y, Han J: **TGF-β signaling and its functional significance in regulating the fate of cranial neural crest cells.** *Crit Rev Oral Biol Med.* 2003; **14**(2): 78–88.
PubMed Abstract | **Publisher Full Text**

93. Puga A, Tomlinson CR, Xia Y: **Ah receptor signals cross-talk with multiple developmental pathways.** *Biochem Pharmacol.* 2005; **69**(2): 199–207.
PubMed Abstract | **Publisher Full Text**

94. Goodale BC, La Du JK, Bisson WH, *et al.*: **AHR2 mutant reveals functional diversity of aryl hydrocarbon receptors in zebrafish.** *PloS One.* 2012; **7**(1): e29346.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

95. Nurminsky D, Magee C, Faverman L, *et al.*: **Regulation of chondrocyte differentiation by actin-severing protein adseverin.** *Dev Biol.* 2007; **302**(2): 427–37.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

96. Vieira FA, Thorne MA, Stueber K, *et al.*: **Comparative analysis of a teleost skeleton transcriptome provides insight into its regulation.** *Gen Comp Endocrinol.* 2013; **191**: 45–58.
PubMed Abstract | **Publisher Full Text**

97. Furuse M, Oda Y, Higashi T, *et al.*: **Lipolysis-stimulated lipoprotein receptor: a novel membrane protein of tricellular tight junctions.** *Ann N Y Acad Sci.* 2012; **1257**: 54–8.
PubMed Abstract | **Publisher Full Text**

98. Jazag A, Ijichi H, Kanai F, *et al.*: **Smad4 silencing in pancreatic cancer cell lines using stable RNA interference and gene expression profiles induced by transforming growth factor-beta.** *Oncogene.* 2005; **24**(4): 662–71.
PubMed Abstract | **Publisher Full Text**

99. Planchart A, Mattingly CJ: **2,3,7,8-Tetrachlorodibenzo-*p*-dioxin upregulates *FoxQ1b* in zebrafish jaw primordium.** *Chem Res Toxicol.* 2010; **23**(3): 480–7.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

100. Hering NA, Andres S, Fromm A, *et al.*: **Transforming growth factor-*β*, a whey protein component, strengthens the intestinal barrier by upregulating claudin-4 in HT-29/B6 cells.** *J Nutr.* 2011; **141**(5): 783–9.
PubMed Abstract | **Publisher Full Text**

101. Ito Y, Yeo JY, Chytil A, *et al.*: **Conditional inactivation of *Tgfbr2* in cranial neural crest causes cleft palate and calvaria defects.** *Development.* 2003; **130**(21): 5269–80.
PubMed Abstract | **Publisher Full Text**

102. Kedersha NL, Miquel MC, Bittner D, *et al.*: **Vaults. II. Ribonucleoprotein structures are highly conserved among higher and lower eukaryotes.** *J Cell Biol.* 1990; **110**(4): 895–901.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

103. Berger W, Steiner E, Grusch M, *et al.*: **Vaults and the major vault protein: novel roles in signal pathway regulation and immunity.** *Cell Mol Life Sci.* 2009; **66**(1): 43–61.
PubMed Abstract | **Publisher Full Text**

104. van Driel M, Pols HA, van Leeuwen JP: **Osteoblast differentiation and control by vitamin D and vitamin D metabolites.** *Curr Pharm Des.* 2004; **10**(21): 2535–55.
PubMed Abstract | **Publisher Full Text**

105. Völkers M, Rohde D, Goodman C, *et al.*: **S100A1: a regulator of striated muscle sarcoplasmic reticulum Ca²⁺ handling, sarcomeric, and mitochondrial function.** *Curr Pharm Des.* 2010; **2010**: 178614.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

106. McBride HM, Neuspiel M, Wasiak S: **Mitochondria: more than just a powerhouse.** *Curr Biol.* 2006; **16**(14): R551–60.
PubMed Abstract | **Publisher Full Text**

107. O'Brien KM, Mueller IA: **The unique mitochondrial form and function of Antarctic channichthyid icefishes.** *Integr Comp Biol.* 2010; **50**(6): 993–1008.
PubMed Abstract | **Publisher Full Text**

108. William J, Ballard O, Pichaud N: **Mitochondrial DNA more than an evolutionary bystander.** *Functional Ecol.* 2014; **28**(1): 218–231.
Publisher Full Text

109. Teacher AG, André C, Merilä J, *et al.*: **Whole mitochondrial genome scan for population structure and selection in the Atlantic herring.** *BMC Evol Biol.* 2012; **12**: 248.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

110. Gudbrandsson J, Ahi E, Franzdottir S, *et al.*: **Dataset 1 in: The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs.** *F1000Research.* 2015.
Data Source

111. Gudbrandsson J, Ahi E, Franzdottir S, *et al.*: **Dataset 2 in: The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs.** *F1000Research.* 2015.
Data Source

112. Gudbrandsson J, Ahi E, Franzdottir S, *et al.*: **Dataset 3 in: The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs.** *F1000Research.* 2015.
Data Source

![F1000Research]

Open Peer Review

Current Referee Status: [?] [?]

Version 1

[?] **Anne Dalziel**

Institute for Systems and Integrative Biology (IBIS), Department of Biology, Laval University, Quebec City, QC, Canada

In this paper "The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs" Gudbrandsson *et al*. have tested for differential gene expression at multiple developmental time-points among a number of Artic charr morpho-types from Lake Thingvallavatn (3 wild morphs, 1 studied with RNA-seq and qPCR, the others with qPCR only) and Holar aquaculture (1 domesticated morph, RNA-seq and qPCR). They have also studied multiple tissues/body regions for a subset of the differentially expressed genes found with RNA-seq. The goal of the paper was to find candidate genes that may underlie variation in morphology, with a focus on craniofacial morphology related to benthic vs. limnetic feeding. In general, I think this goal was met and this paper contributes to our understanding of the mechanisms contributing to morphological evolution in a non-genetic model organism. The authors provide an extensive, multi-time point comparison of two morphologically divergent groups of charr reared in a common environment (reducing the influence of phenotypic plasticity) and have collected a tremendous amount of data. This information will help them to hone in on the genetic loci contributing to phenotypic evolution in this very interesting system, and on the effects of domestication. However, there are a number of major issues that do need to be more clearly addressed in the manuscript prior to final publication. I have outlined these comments below.

Major Comments
1. Introduction:

   Requires some reorganization, clarification of what phenotypes have evolved in parallel among morphs, and how the authors separate the effects of domestication (SB vs. AC) from benthic/limnetic evolution (SB/LB vs. PL/AC).

   a) At present, the introduction focuses upon the utility of instances of parallel evolution to help us determine how repeatable evolutionary change may be. This is definitely true, and the repeated evolution of the dwarf, benthic morph (SB; the focus of the introduction/abstract/discussion) in many lakes strongly argues that this phenotype has evolved via natural selection. However, it is not clear to me if true 'parallelism' seen among the SB (small benthic) and LB (large benthivorous) vs. AC (Holar aquaculture) and PL (small planktivorous) morphs because not enough information is provided for me to assess this. To support the argument for parallelism the specific traits that have evolved in parallel among morphs must be displayed and the evolutionary history of these morphs should be clarified (e.g. in paragraph 6 and Figure 1). As well, any related non-parallelism in traits should also be discussed (i.e. how are the domesticated AC and wild PL different?). At present

Figure 1 only shows the AC and SB morphs, and does not point out the specific traits they are interested in. This is critical background information for readers who are not familiar with this system.

b) The comparison of AC (domestic, limnetic-like head) vs. LB (wild, benthic like head) looks at two confounded variables: domestication and the benthic/limnetic morphology. This should be clearly stated in the introduction, and the use of the additional morphs (PL, LB) in detangling domestication vs. benthic/limnetic evolution should be noted.

c) The use of the AC morph is still a bit unclear to me. The argument for point 'ii) of the availability of abundant AC material' could be expanded by providing more information on the 'limnetic' like features of this morph and why it is an appropriate comparison to a benthic morph, the genetic divergence from the lake Thingvallavatn fish, and also the selection regime it has experienced (selection for limnetic features? What other traits vary with domestication?).

d) Paragraph 2 – Much of this paragraph, including discussing the ability to measure gene expression and relate to phenotype in fishes, is unnecessary as fish are no different from other vertebrates in this respect. Instead, the final sentence "One approach to identify pathways related to function or morphological differences is to study gene expression during development" should become the 'topic sentence' and expanded upon to explain why gene expression studies are especially relevant ways to link genotype to phenotype in evo-devo studies.

e) Better highlight the strengths – The authors have done a wonderful job of assessing multiple developmental time points and rearing fish in a common garden environment. However, they do not highlight these strengths. Some small notes on the importance of controlling for phenotypic plasticity in these traits (which are known to be quite plastic) to better study genetic differentiation would be a nice addition.

2. Methods:

a) Page 4 paragraph 1 - Clarify the number of fish used to make the crosses (this will help us determine the likelihood of selecting a full or half-sib for sequencing/qPCR).

b) I should note that I am not an expert in the analysis of RNA-seq data, but luckily the first reviewer has done an excellent job of commenting upon these aspects of the project. I fully agree with their comments and suggestions. I would also like to see more information on the methods used to pool samples and how RNA-seq data was normalized among samples, developmental times and morphs. I will also note that the authors often use *S.salar* for comparisions, not *O.mykiss*, which is a closer relative to *S.alpinus*. The reasons for this approach should be discussed.

c) I am also not trained as a population geneticist. However, from my experience studying paralogous genes in salmonids, and with respect to the author's own findings for the Nattl paralogs (Fig 4), I do not think it is prudent to "assume that the expression of paralogous genes is stable… " in the methods (page 12). In fact, Berthelot *et al*. (2014) find the opposite (see my comments for the discussion).

d) The authors should use their genetic information to test if the fish chosen are siblings with each other (full or half-sibs). This may have important implications for the population genetic analyses.

e) Page 5 - It is not appropriate to change the meaning of the word 'gene'. I think it is much clearer to use the term 'paralog group' or 'gene family' when referring to the fact that the authors do not study single genes, but instead groups of paralogs.

f) Selection of genes for qPCR – the methods by which genes for the qPCR studies (Fig 3) were selected should be clearly noted. From my reading, it seems that most of these genes do not significantly vary among SB and AC at the 1% FDR level (Tables 1 and 2; only Natterin?). Thus, I am assuming these genes are only significant at the 5% FDR level (S1 file) – why focus upon these and not those significant at 1%? As well, it would be good to include information on why different genes were selected for Figure 3 (qPCR validation of whole fish) and Figure 4 (candidate genes-qPCR validation in just the head). Finally, the abbreviations used for qPCR validation should also be listed in Table 1 for easy comparisons among figures/tables.

3. Results & Figures:

a) Include an experimental design figure - At present, it is difficult to keep track of all of the morphotypes, tissues, and developmental time points used without referring to the methods. Thus, an experimental design figure summarizing the samples used (morphotype, population, sample size, developmental time point), how they were pooled and which techniques were used to measure gene expression on each sample (RNA-seq and/or qPCR) is needed.

b) Include the LB and PL morphs in Figure 1 and clarify traits of interest – The legend states that "differences in size, coloration and head morphology are apparent", but it would be better to specifically point out the differences they are referring to. F1000 is for a general audience, and this would help non-ichthyologists better understand what ecologically-important traits the authors are interested in (e.g. those related to benthic/limnetic feeding). In addition, the two other morphs used in the qPCR studies should also be displayed (large benthivorous and small planktivorous) to facilitate phenotypic comparisons and assess parallelism in benthic/limnetic feeding and/or the effects of domestication on AC.

c) Figure 5- this is actually a table not a figure (?) and is a bit confusing. I think it is much easier to interpret Figure S2 (displaying the data as in Fig 3 and 4), and that Fig 5 and S2 should be switched. It would be great to show significant differences in mRNA content in this, and all other figures, by including symbols. Also, full gene names should be listed in all figure legends.

4. Discussion:

The discussion focuses on the SB morph (page 17 – "The objective of this study were to get a handle on genetic and molecular systems that associate with benthic morphology in charr by mainly focusing on the small benthic morph in Lake Thingvallavatn, Iceland"), while the introduction discusses parallel evolution (indicating that the comparisons should be among many morphs). These are two different topics i) mRNA content differences among benthic vs. limnetic morphs changing in parallel or ii) linking mRNA content to phenotype in SB (benthic, wild) vs. AC (limnetic head, domesticated) morphs. In particular, the role of domestication vs. wild fish divergence needs to be addressed. At present these two topics/questions are mixed in the introduction/discussion and should be addressed separately.

a) Paragraph on Immune Defences - Is immunity also expected to evolve in parallel in all benthic morphs? Is this predicted to be unique to SB vs. AC? Whatever the case, the parallelism (or not) in

these genes should also be discussed, and whether this relates more to domestication in AC or differences between limnetic vs. benthic fish. Much of the functional discussion can also be cut.

b) Page 18 – The information about genes found to be differentially expressed among morphs in your prior work should also be in the introduction, as it is background work that explains why you took this transcriptomic approach. This can also be used to explain why you focused in on particular qPCR genes.

c) A discussion of domestication related differences vs. benthic/limnetic differences should be included. I think the data from head gene expression is very interesting (Figs 5, S2) and really speaks to this question.

d) In general, the role of stochastic evolutionary processes, and not just selection (artificial and natural) should be noted. For example, if the AC charr were simply taken from a stock with a different mtDNA haplotype then these differences in the mtDNA genome might not be adaptive, just random. If the AC fish has much higher mtDNA expression might this be simply a domestication issue and not indicative of selection in SB as stated?   Finally, you find that not all mitochondrial transcripts (which are transcribed as a polycistronic transcript) are found at similar levels (Table 1) – what does this tell you about differential degradation/post-transcriptional processes?

e) There is no discussion about the "Analyses of polymorphism in Arctic charr transcriptome" (Table 3, 4, 5), except for the mtDNA.

Minor Comments

Introduction:

a) Paragraph 3 – "Furthermore, recent estimates from the rainbow trout….by utilizing multiple data sources the genome assembly problem of this family can be solved". I am not sure how this statement is relevant to this particular study. This and the following statement seem more appropriate for the methods/discussion to me.

b) The morphs being discussed should be clarified throughout the paper. For example, the authors often state "among morphs/among charr populations" but it is not clear which of the many morphs they are referring to (e.g. Paragraph 5, first sentence on allozymes and mtDNA and later sentence on MCHIIa – do you mean all 4 morphs of specific 2-way comparisons? Are some morphs more differentiated than others?)

Methods:

a) The authors should note why they did not use the PI (large piscivorous) morph in any qPCR studies (in the methods or discussion) as this would be a nice morph to use in their tests for parallelism.

b) Page 5 (last paragraph) – the methods used to remove particular variants needs to be clarified. In particular, why the assumptions used to remove variants are valid by referencing past studies.

Figures & Results:

a) Figure 2. The key for Figure 2 should include a specific heading for morph and time-point with the abbreviations restated [e.g. Timepoint: 141 dpf, Morph: Small Benthic (SB)].

b) Figure 6 – would be helpful to label the protein coding genes in this figure as well as the 12s and 16s RNAs.

c) Figure 7 – It is not clear to me which variant is present in which morph. Adding the nucleotide to the x-axis (i.e. frequency of m1829G for B) would make this figure easier to quickly interpret. The "A.charr_WT" and "A.charr_M" should also be defined in the legend and it would be more appropriate to use scientific names for all species.

Discussion:

a) Discussion of reference 32 – The discussion of reference 32 is not put into the proper context. Figure 6 of this paper (Berthelot *et al*. 2014) shows that there are many genes that have no correlation among expression patterns and/or differences in expression levels (1573, 1248, and 1895=4716 paralog pairs), and that together these represent more than the 1,407 correlated/similar expression level paralogs. This section of the discussion needs to be modified.

b) The Norman *et al*. (2014) paper should be mentioned earlier – if this is available why was it not used for their analyses? As well, the last sentence in this paragraph can be cut as it is evident.

c) Page 18 – "Our new data also demonstrate differences in craniofacial elements between AC- and SB-charr, along a limnetic vs. benthic axis[79]". Are you referring to ref 79 or data from this study? If you are referring to 79, clarify and note what you found. This occurs a few times in the discussion

General grammatical errors

There are a number of grammatical errors throughout this paper (e.g. "31 genes were higher expressed in SB and 40 genes higher in AC-charr"; "that may help sculpture benthic vs. limnetic heads" pg 19).

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Arnar Palsson, University of Iceland, Iceland

Major Comments

Introduction:

Requires some reorganization, clarification of what phenotypes have evolved in parallel among morphs, and how the authors separate the effects of domestication (SB vs. AC) from benthic/limnetic evolution (SB/LB vs. PL/AC).

a) At present, the introduction focuses upon the utility of instances of parallel evolution to help us

determine how repeatable evolutionary change may be. This is definitely true, and the repeated evolution of the dwarf, benthic morph (SB; the focus of the introduction/abstract/discussion) in many lakes strongly argues that this phenotype has evolved via natural selection. However, it is not clear to me if true 'parallelism' seen among the SB (small benthic) and LB (large benthivorous) vs. AC (Holar aquaculture) and PL (small planktivorous) morphs because not enough information is provided for me to assess this. To support the argument for parallelism the specific traits that have evolved in parallel among morphs must be displayed and the evolutionary history of these morphs should be clarified (e.g. in paragraph 6 and Figure 1). As well, any related non-parallelism in traits should also be discussed (i.e. how are the domesticated AC and wild PL different?). At present Figure 1 only shows the AC and SB morphs, and does not point out the specific traits they are interested in. This is critical background information for readers who are not familiar with this system.

Reply: These are excellent suggestions. At the end of the intro we stress the difference between the aims of our research program (study the genetics of parallel evolution) and the aims of this study (get a handle on differences between sympatric morphs, with the AC as possible outgroup). The morphs studied here do not represent parallel evolution of benthic phenotypes (SB and LB are both from the same lake and appear to be closely related - Kapralova et al 2011). Analyses of that question requires further studies. This data can implicate genes that separate PL/AC and SB/LB and may be studied in such follow up analyses of more populations. We have updated figure 1 as advised - including the 4 morphs studied, expanded on the legend and also provide an overview of research approach (part B).

b) The comparison of AC (domestic, limnetic-like head) vs. LB (wild, benthic like head) looks at two confounded variables: domestication and the benthic/limnetic morphology. This should be clearly stated in the introduction, and the use of the additional morphs (PL, LB) in detangling domestication vs. benthic/limnetic evolution should be noted.

c) The use of the AC morph is still a bit unclear to me. The argument for point 'ii) of the availability of abundant AC material' could be expanded by providing more information on the 'limnetic' like features of this morph and why it is an appropriate comparison to a benthic morph, the genetic divergence from the lake Thingvallavatn fish, and also the selection regime it has experienced (selection for limnetic features? What other traits vary with domestication?).

Reply (b and c): The reviewer is correct, AC and SB are separated by multiple traits, and the data probably reveal signals associating with most of them. Unfortunately the AC charr is not well characterized phenotypically, thus we can not address the question of other traits. We focus mainly on the head and jaw morphology, as these attributes distinguish benthic and limnetic morphs. The revised intro elaborates on the choice of AC, and how the follow up work on the morphs from Lake Thingvallavatn can help us sort this out. This point is also picked up in the discussion.

d) Paragraph 2 – Much of this paragraph, including discussing the ability to measure gene expression and relate to phenotype in fishes, is unnecessary as fish are no different from other vertebrates in this respect. Instead, the final sentence "One approach to identify pathways related to function or morphological differences is to study gene expression during development" should become the 'topic sentence' and expanded upon to explain why gene expression studies are especially relevant ways to link genotype to phenotype in evo-devo studies.

Reply: We restructured and shortened this paragraph around this topic sentence - and gave more

room for the previous RNAseq study on Arctic charr.

e) Better highlight the strengths – The authors have done a wonderful job of assessing multiple developmental time points and rearing fish in a common garden environment. However, they do not highlight these strengths. Some small notes on the importance of controlling for phenotypic plasticity in these traits (which are known to be quite plastic) to better study genetic differentiation would be a nice addition.

Reply: Great advice, we tried to integrate this into the last paragraph of the intro.

Methods:

a) Page 4 paragraph 1 - Clarify the number of fish used to make the crosses (this will help us determine the likelihood of selecting a full or half-sib for sequencing/qPCR).

Reply: We did bulk crosses, joining eggs from 5-10 females in a can and sperm from 3-5 males (SB, PL, LB) and single parent cross for AC. Each sample included RNA pooled from 3 embryos, so there is a chance that full sibs were sequenced, but unlikely. The embryos/samples for qPCR are from similar pools. Now described better in methods.

b) I should note that I am not an expert in the analysis of RNA-seq data, but luckily the first reviewer has done an excellent job of commenting upon these aspects of the project. I fully agree with their comments and suggestions. I would also like to see more information on the methods used to pool samples and how RNA-seq data was normalized among samples, developmental times and morphs. I will also note that the authors often use S.salar for comparisions, not O.mykiss, which is a closer relative to S.alpinus. The reasons for this approach should be discussed.

Reply: The RNA was isolated from individual embryos, quantified and then united (in equal concentrations) prior to cDNA synthesis. The read counts per gene are normalized per million reads in sample. Not normalized with other variables.

c) I am also not trained as a population geneticist. However, from my experience studying paralogous genes in salmonids, and with respect to the author's own findings for the Nattl paralogs (Fig 4), I do not think it is prudent to "assume that the expression of paralogous genes is stable… " in the methods (page 12). In fact, Berthelot et al. (2014) find the opposite (see my comments for the discussion).

Reply: Excellent suggestion. We corrected our misunderstanding, added this fact into the intro and discussion, and reinterpreted our data in this light.

d) The authors should use their genetic information to test if the fish chosen are siblings with each other (full or half-sibs). This may have important implications for the population genetic analyses.

Reply:The fish chosen for pop-gen work are random sample from spawning grounds - assumed to be not sibling groups. Our earlier study (Kapralova 2011) showed no family structure in charr collected this way from the lake.

e) Page 5 - It is not appropriate to change the meaning of the word 'gene'. I think it is much clearer to use the term 'paralog group' or 'gene family' when referring to the fact that the authors do not

study single genes, but instead groups of paralogs.

Reply: Excellent suggestion. We amended this., and use paralog group throughout.

f) Selection of genes for qPCR – the methods by which genes for the qPCR studies (Fig 3) were selected should be clearly noted. From my reading, it seems that most of these genes do not significantly vary among SB and AC at the 1% FDR level (Tables 1 and 2; only Natterin?). Thus, I am assuming these genes are only significant at the 5% FDR level (S1 file) – why focus upon these and not those significant at 1%? As well, it would be good to include information on why different genes were selected for Figure 3 (qPCR validation of whole fish) and Figure 4 (candidate genes-qPCR validation in just the head). Finally, the abbreviations used for qPCR validation should also be listed in Table 1 for easy comparisons among figures/tables.

Reply: Very important point. We deliberately studied some genes with less statistical support (FDR between 5% and 10%), to gauge the differences in the genes with less support and in particular to have a bigger pool of candidates that may relate to the specific developmental process (like head and jaw formation). Of course we can not assert that all the genes with strongest DE signal in the transcriptome are true positives, but the data can be used for hypothesis generation. We also amended table 1 and the figure legends accordingly.

Results & Figures:

a) Include an experimental design figure - At present, it is difficult to keep track of all of the morphotypes, tissues, and developmental time points used without referring to the methods. Thus, an experimental design figure summarizing the samples used (morphotype, population, sample size, developmental time point), how they were pooled and which techniques were used to measure gene expression on each sample (RNA-seq and/or qPCR) is needed.

b) Include the LB and PL morphs in Figure 1 and clarify traits of interest – The legend states that "differences in size, coloration and head morphology are apparent", but it would be better to specifically point out the differences they are referring to. F1000 is for a general audience, and this would help non-ichthylogists better understand what ecologically-important traits the authors are interested in (e.g. those related to benthic/limnetic feeding). In addition, the two other morphs used in the qPCR studies should also be displayed (large benthivorous and small planktivorous) to facilitate phenotypic comparisons and assess parallelism in benthic/limnetic feeding and/or the effects of domestication on AC.

Reply: (a and b) Excellent suggestions. Now picture 1 has all 4 morphs, and a schematic describing the work flow and samples.

c) Figure 5- this is actually a table not a figure (?) and is a bit confusing. I think it is much easier to interpret Figure S2 (displaying the data as in Fig 3 and 4), and that Fig 5 and S2 should be switched. It would be great to show significant differences in mRNA content in this, and all other figures, by including symbols. Also, full gene names should be listed in all figure legends.

Reply: We acknowledge that this graph is not the simplest, but would like to keep it over Figure S2. Our reasoning is that this graph illustrates the sharp differences between the limnetic (AC-PL) and benthic (SB-LB), which are the main result in this section. But we will of course switch them, or possibly join both in a single figure ?? if the reviewer insists or the editors recommend it.

Discussion:

The discussion focuses on the SB morph (page 17 – "The objective of this study were to get a handle on genetic and molecular systems that associate with benthic morphology in charr by mainly focusing on the small benthic morph in Lake Thingvallavatn, Iceland"), while the introduction discusses parallel evolution (indicating that the comparisons should be among many morphs). These are two different topics i) mRNA content differences among benthic vs. limnetic morphs changing in parallel or ii) linking mRNA content to phenotype in SB (benthic, wild) vs. AC (limnetic head, domesticated) morphs. In particular, the role of domestication vs. wild fish divergence needs to be addressed. At present these two topics/questions are mixed in the introduction/discussion and should be addressed separately.

Reply: We tried to separate these two aims more clearly in the revised discussion. The strategy was to use the AC vs SB contrast for hypothesis generation, as the first aim is central to our program. We have now added sentences on the domestication in two parts of the discussion.

a) Paragraph on Immune Defenses - Is immunity also expected to evolve in parallel in all benthic morphs? Is this predicted to be unique to SB vs. AC? Whatever the case, the parallelism (or not) in these genes should also be discussed, and whether this relates more to domestication in AC or differences between limnetic vs. benthic fish. Much of the functional discussion can also be cut.

Reply: Good question, we assume it to be so, but that may be wrong. We moved the discussion towards this question and away from functional description.

b) Page 18 – The information about genes found to be differentially expressed among morphs in your prior work should also be in the introduction, as it is background work that explains why you took this transcriptomic approach. This can also be used to explain why you focused in on particular qPCR genes.

Reply: We added a sentence in the intro about the published papers, that this transcriptome made available. In those papers we focused on genes with putative craniofacial effects, though the focus in this study was broader.

c) A discussion of domestication related differences vs. benthic/limnetic differences should be included. I think the data from head gene expression is very interesting (Figs 5, S2) and really speaks to this question.

d) In general, the role of stochastic evolutionary processes, and not just selection (artificial and natural) should be noted. For example, if the AC charr were simply taken from a stock with a different mtDNA haplotype then these differences in the mtDNA genome might not be adaptive, just random. If the AC fish has much higher mtDNA expression might this be simply a domestication issue and not indicative of selection in SB as stated? Finally, you find that not all mitochondrial transcripts (which are transcribed as a polycistronic transcript) are found at similar levels (Table 1) – what does this tell you about differential degradation/post-transcriptional processes?

e) There is no discussion about the "Analyses of polymorphism in Arctic charr transcriptome" (Table 3, 4, 5), except for the mtDNA.

Reply: (c,d,e) Excellent suggestions. We added in the final discussion section few sentences on domesticated charr vs Benthic/limnetic. Unfortunately we do not have quantitative data on the phenotypes (head shape, and jaw) of the AC charr and acknowledge that we categorize it as limnetic based on general features.

We gladly added a sentence citing neutral forces, and are acutely aware that much of the divergence is likely due to history, drift etc. The domestication can certainly be the driver for the higher expression in AC - but we need transcriptomes from more populations/morphs to address that point. And yes, the variance in RNA levels from different parts of the mtDNA do indeed suggest differential half life of the various RNA species. Some are certainly degraded and others most probably actively utilized / protected. We decided not to follow that thought further though, as the MS already consists of quite a few threads already.

We also added sentences on the genetic polymorphism, before focusing more on the mtDNA. The main reason we dont want to elaborate to much on the SNPs is that we feel these data are mainly for generating hypotheses, and that more work is needed to substantiate SNPs and study their distribution in other populations.

Minor Comments

Introduction:

a) Paragraph 3 – "Furthermore, recent estimates from the rainbow trout….by utilizing multiple data sources the genome assembly problem of this family can be solved". I am not sure how this statement is relevant to this particular study. This and the following statement seem more appropriate for the methods/discussion to me.

Reply: We deleted this sentence and simplified the paragraph.

b) The morphs being discussed should be clarified throughout the paper. For example, the authors often state "among morphs/among charr populations" but it is not clear which of the many morphs they are referring to (e.g. Paragraph 5, first sentence on allozymes and mtDNA and later sentence on MCHIIa – do you mean all 4 morphs of specific 2-way comparisons? Are some morphs more differentiated than others?)

Reply: We tried to clarify this in various places in the manuscript, but in some cases we refer to morphs in general. Genetic separation can be estimated with Fst values either between pairs or over a larger set of groups (populations, morphs). In the intro we cite the work done to date in Iceland, which highlights the need for more pop. genetic analyses.

Methods:

a) The authors should note why they did not use the PI (large piscivorous) morph in any qPCR studies (in the methods or discussion) as this would be a nice morph to use in their tests for parallelism.

Reply: The PI charr is very rare in the lake and hard to catch. We later captured few sexually mature individuals, and generated couple of families, that were used for one study (Ahi et al

Evodevo 2015).

b) Page 5 (last paragraph) – the methods used to remove particular variants needs to be clarified. In particular, why the assumptions used to remove variants are valid by referencing past studies.

Reply: Many of the principles are common to most pipelines for removing spurious variants. In addition we applied filters necessitated by the properties of our dataset (pool of individuals), the mapping to an outgroup and paralogs due to salmonid genome complexity.

Figures & Results:

a) Figure 2. The key for Figure 2 should include a specific heading for morph and time-point with the abbreviations restated [e.g. Timepoint: 141 dpf, Morph: Small Benthic (SB)].
Reply: Now fixed.

b) Figure 6 – would be helpful to label the protein coding genes in this figure as well as the 12s and 16s RNAs.
Reply: Now fixed.

c) Figure 7 – It is not clear to me which variant is present in which morph. Adding the nucleotide to the x-axis (i.e. frequency of m1829G for B) would make this figure easier to quickly interpret. The "A.charr_WT" and "A.charr_M" should also be defined in the legend and it would be more appropriate to use scientific names for all species.
Reply: Now fixed

Discussion:

a) Discussion of reference 32 – The discussion of reference 32 is not put into the proper context. Figure 6 of this paper (Berthelot et al. 2014) shows that there are many genes that have no correlation among expression patterns and/or differences in expression levels (1573, 1248, and 1895=4716 paralog pairs), and that together these represent more than the 1,407 correlated/similar expression level paralogs. This section of the discussion needs to be modified.

Reply: Really valuable point, that we are especially grateful for. That we have added this fact to the intro and altered our interpretations in the discussion.

b) The Norman et al. (2014) paper should be mentioned earlier – if this is available why was it not used for their analyses? As well, the last sentence in this paragraph can be cut as it is evident.

Reply: The Norman papers are now presented more clearly in the intro. There are historical reasons for not including their data in our analyses, we had completed the analyses for this manuscript when they became available and have since then focused our data analyses efforts on another transcriptome generated in the lab (with longer reads).

c) Page 18 – "Our new data also demonstrate differences in craniofacial elements between AC- and SB-charr, along a limnetic vs. benthic axis79". Are you referring to ref 79 or data from this study? If you are referring to 79, clarify and note what you found. This occurs a few times in the discussion

Reply: Ref 79 is a related study that built in part on the data presented here. We have now rephrased this in the manuscript, hopefully to the better.

Competing Interests: No competing interests were disclosed.No competing interests were disclosed.

Referee Report 07 July 2015

doi:10.5256/f1000research.6869.r8970

Daniel Macqueen
Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, UK

Review of Gudbrandsson *et al*. "The developmental transcriptome of contrasting Arctic charr (*Salvelinus alpinus*) morphs".

The work is founded on the solid premise that rapidly evolving phenotypes in nature can be underpinned by changes at the transcriptome level. The model system here is Arctic charr populations that have evolved (since the last ice age) major differences in phenotypes along the 'benthic' - 'limnetic' axis, with strong differences in head morphology linked to feeding specializations. The work provides an extensive analysis of transcriptome and genetic differences between different morphs and populations. It is interesting, generally well-written and has merit on many levels. It is also rather hard going, since so much ground is covered on diverse areas. The study also comes with a large number of caveats, of which the authors are undoubtedly aware. Overall though, I am supportive of this work, as it represents one of the most detailed analyses of molecular mechanisms linked to rapid phenotypic evolution in Arctic charr. I see it as a great start point for future work and a source of several new findings and hypotheses. I suggest that the paper be indexed in F1000 Research as long as its caveats are transparent and the authors address my comments.

I list below a number of suggestions that may help the authors improve the work, or that at least highlight study limitations for the benefit of interested readers. I also provide a number of minor comments and suggestions, which should help improve the manuscript more incrementally.

Main comments & caveats

1. **RNAseq study design.** I sympathize with the fact that the authors are trying to publish Illumina data that was generated in 2009, since (obviously) the technology has moved on greatly in the last 6 years, while its costs have been reduced dramatically. Adding to this is the fact that the authors are using a particularly complex transcriptome in terms of high content of similar paralogues (and expressed transposable elements), without a reference sequence for mapping in their species. I accept the author's argument that it is more sensible to map against a closely related species with the sequence data rather than to try and create a *de novo* assembly from 36bp reads. I also believe it is sensible to pool read counts for putative paralogous contigs in this study, since the short read length ablates any ability to separate paralogous differences in expression (yet does not preclude the generation of useful hypotheses about putative gene expression differences among morphs).

   However, I do question whether the use of Atlantic salmon EST contigs is the best approach here. Firstly, reference assemblies for both Atlantic salmon and rainbow trout are now available, which

distinguish paralogous variation. More importantly, using these reference genome data would provide certainty that reads are being mapped to exons from single genes, whereas many of the ESTs will provide a fragmented representation of exon sequences, presumably relying on annotation to piece them back into 'genes' *post hoc* . In addition, paired 100bp Ilumina reads are available at high coverage for Arctic charr (e.g. Norman *et al*. 2014), which could also be used to generate a specific reference transcriptome to map against in this study, although this might be underrepresented in terms of developmental genes as it is a gill study. Overall, I do wonder how much more information might have been gleaned from this dataset with a different mapping strategy?

With all the above said, I understand that the authors have built up a large study based around the original mapping to the salmon ESTs and that it would not be routine for them to repeat the study using better reference data. Furthermore, the approach used has definitely led to the generation of several valid hypotheses concerning the nature of gene expression and genetic differences among charr morphs, which have been followed up using independent approaches.

2. **Methods "Biological Replication in RNAseq" –** a general comment: obviously the design of the study is not optimal because biological variation within developmental stages is not considered in the statistics. Thus, the approach lacks power to detect differences when morph variation is restricted to different developmental stages. I wanted to explain my opinion (for the record) that the study design is nonetheless useful for identifying constitutive differences between morphs. This is especially true because gene expression variability is likely to be relatively low in embryonic stages (compared to a similar study design in adults at least). Further, the pooling of individuals will have helped to at least recapture some biological variation at different stages. Thus, as mentioned above, I see the author's use of RNAseq as a hypothesis-generating approach, which has been quite fruitful in identifying putative differences between different morphs.

3. **Methods "QPCR study design".** The authors adhere to the MIQE guidelines, but do not always follow the best approaches. Most pertinently, the authors use the $2^{-\Delta\Delta Ct}$ method (assuming PCR efficiency of 2.0) despite having gone to the effort of gaining and reporting efficiencies for each assay, which can be as low as 1.72 for some genes. The effect of failing to incorporate differences in efficiency are highly established and this is likely to have affected the author's results. The authors should consider incorporating the effect of differences in efficiency into their analyses. This is likely to have some impact on the study conclusions in my opinion.

4. **Methods** "**Polymorphisms in charr transcriptome".** While this is not exactly my area of expertise, I struggled to understand the methods behind filtering paralogous variants from SNPs in the data. The authors state "*As the SNP analysis was done on individual contigs, differences among paralogs appear in the data. However, since each sample is a pool of few individuals, it is very unlikely that we have the same frequency of true SNPs in the samples. This property was used to remove variants that are most likely due to expressed paralogs*". Can the authors please try to re-explain this in even simpler terms to help me get it? I don't see how this description leads to a robust identification of paralogous variation. Is there an underlying assumption of equal expression among paralogues? If so, this is likely to be routinely invalidated.

5. **Methods** "**Verification of candidate SNPs".** While it is good that the authors have attempted to verify SNPs identified from their RNAseq data, I don't believe the data is particularly well incorporated in the results section. It needs to be stated up front the extent to which the SNPs predicted from the RNAseq were independently verified. Also, the methods for this section can be

improved, especially "*we conducted genomic comparisons of the Salmon genome, ESTs and short contigs from the preliminary assembly of the Arctic charr transcriptome*". None of this information is elaborated on – what is the preliminary assembly of the Arctic charr transcriptome? Which version of the salmon genome was used and how? Moreover, it would be useful to actually explain in the methods that the genotyping was done on a small number of SB, PL and PI morphs, rather than relying on the reader to extract all the required information from Table S2. I guess overall, the way this section is incorporated into the manuscript needs some thought in terms of improving the reader's experience. I struggled after reading it several times and am still not sure I have all the information I need.

6. **Results**. "*Analyses of those reads require an Arctic charr genome sequence or transcriptome assembly from longer and paired end reads.*" As mentioned already, the latter is available to generate an Arctic charr transcriptome assembly to map against.

7. **Results**; Figure 3 and 4. The authors found that around half the genes studied were not differentially expressed among morphs by qPCR. Obviously this is quite a large number, but on closer inspection, I noticed that *Ndub6*, *Ubl5* and *parp6* were not even differentially expressed according to RNAseq. Thus, I am confused at the selection of genes from the RNAseq analysis for verification by qPCR. The authors should explain this selection more transparently and provide clearer indices of the correlation between RNAseq and qPCR results and associated discussion.

Minor comments, typos and suggested changes
1. Abstract: "*Species and populations with parallel evolution of specific traits can help illuminate how predictable adaptations and divergence are at the molecular and developmental level*. Grammatically – his reads better: "*….. can help illuminate the predictability of adaptations and divergence at the molecular and developmental level*"

2. Introduction: "*Examples of such a species complex are the finches of the Galapagos islands, cichlids in the African great lakes are exciting multi-species systems in this respect*". Grammatically – reads better: "*Examples of such species complexes are provided by finches of the Galapagos islands, while cichlids of the African great lakes also provide an exciting multi-species system in the same respect*"

3. Introduction: "*Some northern freshwater fish species exhibit frequent parallelism in trophic structures and life history and in several cases are they found as distinct resource morphs*" change to "*…. are found as distinct resource morphs*"

4. Introduction: "*in the development of ecological differences in tropic morphology*" change to "*… trophic morphology*".

5. Introduction: "*The family is estimated to be between 63.2 and 58.1 million years old*". This information is not correct – it is correct to state that the age of the salmonid crown (based on the cited paper; different estimates exist in the literature, e.g. Macqueen and Johnston, 2014; Campbell *et al*. 2013) is estimated at 63.2 and 58.1 million years old, but the family dates back much further – to the origin of the WGD event in fact, which occurred more like 88-103 Ma (Macqueen and Johnston, 2014; Berthelot *et al*. 2014). Thus, the last common ancestor to extant salmonid species is what the authors are actually referring to in this sentence.

6. Introduction: "*Furthermore, for data with short reads, mapping to a related reference genome/transcriptome is recommended over de novo assembly*". While this sentence is technically correct in the context of the work cited, I feel it is being used slightly out of context. For a start, what comprises a 'short read' is undefined. 36bp is short, but it is possible to get a sold reference transcriptome using 2*100bp, assuming the appropriate diversity of transcripts is represented and suitable depth is attained.

7. Introduction: "*nuclear genes, reveled both subtle*" change to "*nuclear genes, revealed both subtle*"

8. Minor comment – AC, PL, LB and SB were already defined in introduction.

9. Methods: "*Fishing in Lake Thingvallavatn was with permissions*" changed to "*Fishing in Lake Thingvallavatn was done with permissions*".

10. Methods: "*of differently expressed genes, we preformed clustering analyses*" change to "*…we performed clustering analyses*"

11. Results: "*The most drastic changes were seen in processes related to glycolysis (GO:0006096, FDR = 0.0009), were the expression of 19 out of 25 genes*" change to "*…. where the expression*".

12. Figure 7. What does the charr_WT vs. charr_M signify in the alignment data?

13. Discussion "*We are interested in how predictable evolution is a the molecular level and if there certain principles influence the rewiring of developmental and regulatory systems during evolution*" consider changing to "*We are interested in the predictability of evolution at the molecular level, especially whether there exist principles that influence the rewiring of developmental and regulatory systems*".

14. Discussion. "*Recent rainbow trout data shows most paralogs from the latest whole genome duplication event retain the same expression pattern[32] indicating that this scenario is probably uncommon; hence it is of considerable interest when two paralogs show distinct expression patterns*". I do not agree that it is of considerable interest when two paralogs show distinct expression patterns – I could list tens of examples for salmonids.

15. Conclusions "*The results suggest genetic and expression changes in multiple systems relate to divergence among populations*." Change to "*… associated with divergence among populations*."

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 04 Apr 2016
Arnar Palsson, University of Iceland, Iceland

Main comments & caveats

RNAseq study design. I sympathize with the fact that the authors are trying to publish Illumina

data that was generated in 2009, since (obviously) the technology has moved on greatly in the last 6 years, while its costs have been reduced dramatically. Adding to this is the fact that the authors are using a particularly complex transcriptome in terms of high content of similar paralogues (and expressed transposable elements), without a reference sequence for mapping in their species. I accept the author's argument that it is more sensible to map against a closely related species with the sequence data rather than to try and create a de novo assembly from 36bp reads. I also believe it is sensible to pool read counts for putative paralogous contigs in this study, since the short read length ablates any ability to separate paralogous differences in expression (yet does not preclude the generation of useful hypotheses about putative gene expression differences among morphs).

However, I do question whether the use of Atlantic salmon EST contigs is the best approach here. Firstly, reference assemblies for both Atlantic salmon and rainbow trout are now available, which distinguish paralogous variation. More importantly, using these reference genome data would provide certainty that reads are being mapped to exons from single genes, whereas many of the ESTs will provide a fragmented representation of exon sequences, presumably relying on annotation to piece them back into 'genes' post hoc . In addition, paired 100bp Ilumina reads are available at high coverage for Arctic charr (e.g. Norman et al. 2014), which could also be used to generate a specific reference transcriptome to map against in this study, although this might be underrepresented in terms of developmental genes as it is a gill study. Overall, I do wonder how much more information might have been gleaned from this dataset with a different mapping strategy?

With all the above said, I understand that the authors have built up a large study based around the original mapping to the salmon ESTs and that it would not be routine for them to repeat the study using better reference data. Furthermore, the approach used has definitely led to the generation of several valid hypotheses concerning the nature of gene expression and genetic differences among charr morphs, which have been followed up using independent approaches.

Reply: We thank the reviewer for excellent diagnosis and suggestions. The paper describes the (in our humble opinion) most sensible summary of the data, as the writing of the paper started 2 years ago. We did map on the O.mykiss cDNA collection also, got similar results, but opted for reporting on the salmon data to avoid further extending an already long manuscript. We are currently analyzing DE and SNPs on a new assembly (100 bp PE reads - 48 samples - 3 morphs - development), and may include a remapping of this dataset in that.

Methods "Biological Replication in RNAseq" – a general comment: obviously the design of the study is not optimal because biological variation within developmental stages is not considered in the statistics. Thus, the approach lacks power to detect differences when morph variation is restricted to different developmental stages. I wanted to explain my opinion (for the record) that the study design is nonetheless useful for identifying constitutive differences between morphs. This is especially true because gene expression variability is likely to be relatively low in embryonic stages (compared to a similar study design in adults at least). Further, the pooling of individuals will have helped to at least recapture some biological variation at different stages. Thus, as mentioned above, I see the author's use of RNAseq as a hypothesis-generating approach, which has been quite fruitful in identifying putative differences between different morphs.

Reply: We appreciate the reviewers careful analyses of the study and approach. We tried to emphasize the "hypothesis-generation" aspect during the rewrite.

Methods "QPCR study design". The authors adhere to the MIQE guidelines, but do not always follow the best approaches. Most pertinently, the authors use the 2−ΔΔCt method (assuming PCR efficiency of 2.0) despite having gone to the effort of gaining and reporting efficiencies for each assay, which can be as low as 1.72 for some genes. The effect of failing to incorporate differences in efficiency are highly established and this is likely to have affected the author's results. The authors should consider incorporating the effect of differences in efficiency into their analyses. This is likely to have some impact on the study conclusions in my opinion.

Reply: Great point. The qPCR primer efficiencies more than 1.90 can be easily assumed as 2 because of the negligible effects. Since we used LinReg software for efficiencies not the traditional method, it takes into account the efficiencies for each test for a given primer pair and discard those have different and lower efficiencies. However, the Natterin-like paralogues were below the cut-off. The statistical analyses were done on deltaCt values, prior to transformation based on efficiencies used for visualization. We now report the graphs of their expression adjusting for the lower efficiency, and state in the results "Note however, the efficiency of the primers for the nattl genes ranged from 1.72 to 1.77, which suggests this data should be interpreted with caution."

Methods "Polymorphisms in charr transcriptome". While this is not exactly my area of expertise, I struggled to understand the methods behind filtering paralogous variants from SNPs in the data. The authors state "As the SNP analysis was done on individual contigs, differences among paralogs appear in the data. However, since each sample is a pool of few individuals, it is very unlikely that we have the same frequency of true SNPs in the samples. This property was used to remove variants that are most likely due to expressed paralogs". Can the authors please try to re-explain this in even simpler terms to help me get it? I don't see how this description leads to a robust identification of paralogous variation. Is there an underlying assumption of equal expression among paralogues? If so, this is likely to be routinely invalidated.

Reply: We acknowledge this part is a hard read. We rewrote this part of the methods. Here is another summary. Reads from regions that are very similar in paralogous genes can map to both of them. Because we consider also reads that map to many contigs, some of the candidate variants will reflect sequence differences between paralogs, not polymorphism in either paralog. Next we deploy the population genetic argument, since we are sequencing RNA from 6 chromosomes in each sample, then it is very unlikely that a TRUE SNP will be at the same frequency in all of the 8 samples. But variants - that are due to differences bwn paralogs - are likely to be similar in frequency because they are unaffected by the population sampling. This filter is designed to toss those out.

To emphasize the objective is not to find differences between paralogs, but rather to enrich for true SNPs. This method will toss out many sites separating paralogous genes (but not all because some paralogous genes are differentially expressed between morphs or time points).

Methods "Verification of candidate SNPs". While it is good that the authors have attempted to verify SNPs identified from their RNAseq data, I don't believe the data is particularly well incorporated in the results section. It needs to be stated up front the extent to which the SNPs predicted from the RNAseq were independently verified. Also, the methods for this section can be improved, especially "we conducted genomic comparisons of the Salmon genome, ESTs and short contigs from the preliminary assembly of the Arctic charr transcriptome". None of this information is elaborated on – what is the preliminary assembly of the Arctic charr transcriptome? Which version of the salmon genome was used and how? Moreover, it would be useful to actually explain in the

methods that the genotyping was done on a small number of SB, PL and PI morphs, rather than relying on the reader to extract all the required information from Table S2. I guess overall, the way this section is incorporated into the manuscript needs some thought in terms of improving the reader's experience. I struggled after reading it several times and am still not sure I have all the information I need.

Reply: We fixed the methods section to accommodate both reviewers which brought up similar points. We highlight the sampling (8 individuals of 3 morphs), and extend the description of the genomic comparisons. We also extend the discussion of those results.

Results. "Analyses of those reads require an Arctic charr genome sequence or transcriptome assembly from longer and paired end reads." As mentioned already, the latter is available to generate an Arctic charr transcriptome assembly to map against.

Reply: Unfortunately the great Norman *et al.* 2014 data (http://www.ncbi.nlm.nih.gov/pubmed/24368751) came to our attention after we had done these analyses, and started working on our new data (see above). Thus we opted for not redoing the whole analyses for this manuscript, but focus on the verification - and of course working on a new assembly using longer reads.

Results; Figure 3 and 4. The authors found that around half the genes studied were not differentially expressed among morphs by qPCR. Obviously this is quite a large number, but on closer inspection, I noticed that Ndub6, Ubl5 and parp6 were not even differentially expressed according to RNAseq. Thus, I am confused at the selection of genes from the RNAseq analysis for verification by qPCR. The authors should explain this selection more transparently and provide clearer indices of the correlation between RNAseq and qPCR results and associated discussion.

Reply: This reflects the history of the project, and the difference between the preliminary and final analyses. We decided to report on all the data - but explain better in the manuscript the classification of genes tested with qPCR, at 1%, 5% and 10% FDR. In summary, some of the genes tested were above 5% and one even just above 10% FDR. Some of those were not corroborated by qPCR. The number of genes is insufficient to do a statistical comparison of the verification rate at the different FDR levels. A table (new Table 3) - supported with few sentences in the results, hopefully clarifies this.

Minor comments, typos and suggested changes

Abstract: "Species and populations with parallel evolution of specific traits can help illuminate how predictable adaptations and divergence are at the molecular and developmental level. Grammatically – his reads better: "..... can help illuminate the predictability of adaptations and divergence at the molecular and developmental level"
Reply: Thanks - fixed.

Introduction: "Examples of such a species complex are the finches of the Galapagos islands, cichlids in the African great lakes are exciting multi-species systems in this respect". Grammatically – reads better: "Examples of such species complexes are provided by finches of the Galapagos islands, while cichlids of the African great lakes also provide an exciting multi-species system in the same respect"
Reply: Thanks - fixed.

Introduction: "Some northern freshwater fish species exhibit frequent parallelism in trophic structures and life history and in several cases are they found as distinct resource morphs" change to "…. are found as distinct resource morphs"
Reply: Thanks - fixed.

Introduction: "in the development of ecological differences in tropic morphology" change to "… trophic morphology".
 Reply: Thanks - fixed.

Introduction: "The family is estimated to be between 63.2 and 58.1 million years old". This information is not correct – it is correct to state that the age of the salmonid crown (based on the cited paper; different estimates exist in the literature, e.g. Macqueen and Johnston, 2014; Campbell et al. 2013) is estimated at 63.2 and 58.1 million years old, but the family dates back much further – to the origin of the WGD event in fact, which occurred more like 88-103 Ma (Macqueen and Johnston, 2014; Berthelot et al. 2014). Thus, the last common ancestor to extant salmonid species is what the authors are actually referring to in this sentence.

Reply: Thanks so for pointing this out. We changed the text to "local adaptation has been extensively studied in the salmonid family, to which Arctic charr belongs {Fraser2011}. The family is estimated to be between 88-103 million years old {Macqueen2014,Berthelot2014c}. A whole genome duplication event occurred before the radiation of the salmonid family {Davidson2010,Moghadam2011,Macqueen2014,Berthelot2014c} which has provided time for divergence of ohnologous genes (paralogous genes originated by whole genome duplication event). "

Introduction: "Furthermore, for data with short reads, mapping to a related reference genome/transcriptome is recommended over de novo assembly". While this sentence is technically correct in the context of the work cited, I feel it is being used slightly out of context. For a start, what comprises a 'short read' is undefined. 36bp is short, but it is possible to get a sold reference transcriptome using 2*100bp, assuming the appropriate diversity of transcripts is represented and suitable depth is attained.

Reply: Great point, we opted for keeping the point (at this place in the ms) but changing the wording to: In this study we opted to map the reads (36 bp) to a related reference genome/transcriptome {Vijay2013a}, instead of conducting de novo assembly.

Introduction: "nuclear genes, reveled both subtle" change to "nuclear genes, revealed both subtle"
Reply: Thanks fixed.

Minor comment – AC, PL, LB and SB were already defined in introduction.
Reply: Thanks, removed this.

Methods: "Fishing in Lake Thingvallavatn was with permissions" changed to "Fishing in Lake Thingvallavatn was done with permissions".
Reply: Ammended.

Methods: "of differently expressed genes, we preformed clustering analyses" change to "…we

performed clustering analyses"
Reply: Thanks, fixed.

Results: *"The most drastic changes were seen in processes related to glycolysis (GO:0006096, FDR = 0.0009), were the expression of 19 out of 25 genes"* change to *".... where the expression".*
Reply: Thanks, fixed.

Figure 7. What does the charr_WT vs. charr_M signify in the alignment data?
Reply: Designates the two alleles, the legend now makes this explicit.

Discussion "We are interested in how predictable evolution is a the molecular level and if there certain principles influence the rewiring of developmental and regulatory systems during evolution" consider changing to "We are interested in the predictability of evolution at the molecular level, especially whether there exist principles that influence the rewiring of developmental and regulatory systems".
Reply: Thanks, excellent suggestion, included

Discussion. "Recent rainbow trout data shows most paralogs from the latest whole genome duplication event retain the same expression pattern32 indicating that this scenario is probably uncommon; hence it is of considerable interest when two paralogs show distinct expression patterns". I do not agree that it is of considerable interest when two paralogs show distinct expression patterns – I could list tens of examples for salmonids.
Reply: Good point, we have revisited this interpretation (see also point by rev. 1).

Conclusions "The results suggest genetic and expression changes in multiple systems relate to divergence among populations." Change to "… associated with divergence among populations."
Reply: Thanks, fixed.

Competing Interests: No competing interests were disclosed.No competing interests were disclosed.