Genetic variation in different morphs of Arctic Charr

Name: Sara Eiríksdóttir Andersen

Genetic variation in different morphs of Arctic Charr

Sara Eiríksdóttir Andersen

12 ECTS thesis
Research Project

Supervisor
Arnar Pálsson

Faculty of Biology
School of Engineering and Natural Sciences
University of Iceland
Reykjavík, February 2015

Preface

The work presented in this paper is a project concerning genetic variation in different morphs of Arctic Charr. The Arctic Charr (*Salvelinus alpinus*) has four morphtypes which are separated by genetic and morphological differences. When comparing RNA sequencing data to the Salmon genome, we can reveal single nucleotide polymorphisms (called SNPs), some of which differ in frequency between the morphs. These RNA data and SNPs derived from them are the foundation of this work, as they are what diverse the morphs, and also very interesting in the development of life in general.

My supervisor in this project is Arnar Pálsson from Department of Life and Environmental Sciences, University of Iceland. The project started in the end of August 2014 and was concluded January 2015.

Abstract

Lake Thingvallavatn is about 10,000 years old, thus a geographically young lake. The lake is inhabitated by four morphs of the Arctic Charr (*Salvelinus alpinus)*, which was colonized by Anadromous charr. These types have evolved from the same forefather (or "forefish") in about 10,000 years. Not only does this make this lake very special, but Lake Thingvallavatn is also one of the few places where so many morphs of the Arctic Charr exist (1). These four morphs differ in morphology. Analyses of gene expression with RNA-seequencing have also revealed expression differences during development between limnetic and benthic morphotypes (2), and also genetic differences consisting of SNPs. Investigating the genetic and expression differences between the morphs, gives an insight in the divergence and possibly speciation of the morphotypes. This project investigates fixed SNPs in 12 genes, chosen after specific criteria, and also how the SNPs are dispersed within genes, all this via bioinformatics methods. Through experimental work I attempted to genotype SNPs in 12 genes, and could confirm 2 of them in in the four morphs. The data revealed allele frequencies similar to the frequencies estimated from the transcriptome.

Ágrip

Þingvallavatn er jarðfræðilega ungt vatn, um 10.000 ára gamalt. Í vatnið gekk snemma sjóbleikja (*Salvelinus alpinus)*, sem hefur nú skipst í fjögur afbrigði bleikju. Afbrigðin fjögur hafa þróast frá sama forföður á þessum 10.000 árum. Hröð afbrigðamyndun er sérstök, ein einstakt er hversu mörg afbrigði finnast innan vatnsins. Afbrigðin fjögur eru ólík í útliti, stærð og lífsháttum. Rannsóknir á genatjáningu með RNA-raðgreiningar aðferðum sýna mun á genatjáningu í fóstrum sviflægra og botnlægra afbrigða. Einnig sést munur á erfðasamsetningu, sem afhjúpast þegar stakar basabreytingar (single nucleotide polymorphisms: SNPs) eru greindar. Með því að greina erfðafræði og tjáningarmun á milli afbrgiða er hægt að rannsaka mun á afbrigðum og jafnvel tilurð tegunda. Í þessu verkefni voru breytileg set í 12 genum rannsökuð. Einnig var kannað hvernig breytingar dreifast innan gena með mörgum breytingum. Með tilraunum var reynt að staðfesta 12 breytileg set, og voru tvö þeirra staðfest í fjórum afbrigðum. Gögnin benda til að tíðni breytileika sé áþekk í stofna sýni og í umritunarmenginu.

Table of Contents

Introduction

Around 10,000 years ago, Lake Thingvallavatn formed, and shortly thereafter the first fish settled in the lake. One of these was the salmonid fish, Arctic Charr (*Salvelinus alpinus*), which has now, 10,000 years later evolved in to distinct four morphs of the original ancestor The four different morphs are different in size, shape and ecology (3). They are also genetically distinct, which can be seen by numerous polymorphic positions in their genes, see Appendix A. When looking in some genes, the frequencies of SNPs are not equally distributed between the morphs. These SNPs might indicate that some of them associate with expression of certain genes coding for a specific effect in the morph.

Theory

The four morphs of Arctic Charr

The Arctic Charr is a cold-water fish with habitat in subarctic and northern temperate freshwater systems, amongst these Lake Thingvallavatn, Icelands largest lake. In Lake Thingvallavatn there exist 4 morphs of the Arctic Charr, which all have evolved from the same fish during the last 10,000 years (4, 5). Two morphs are benthivorous (large benthic, LB and small benthic, SB), and the two other are planktivorous (PL) and piscivorous (PI) (4).

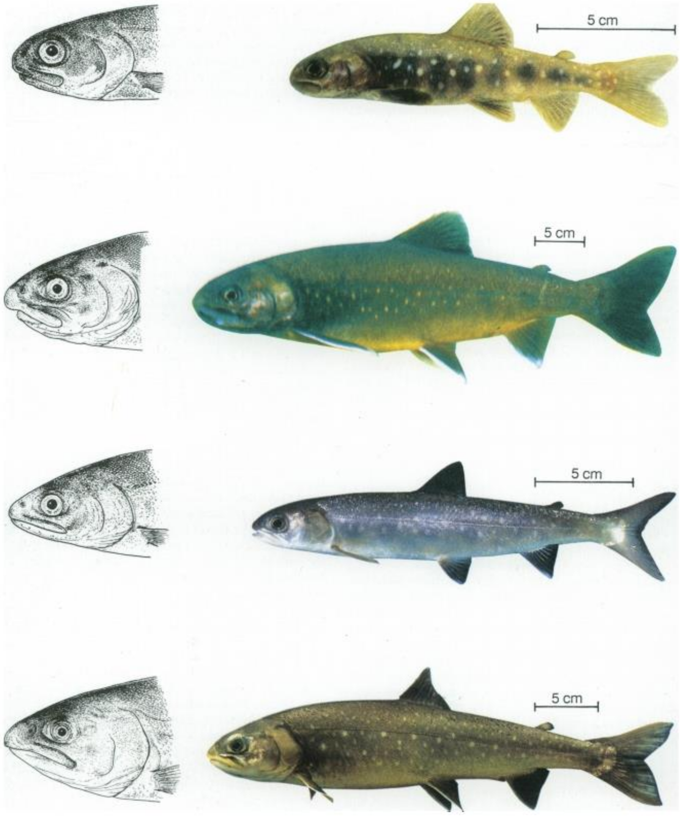


Figure 1: The four morphotypes of Arctic Charr fish, reprinted from (4).

The four morphs differ in several ways, food, colours, size, and use of habitat in lake (4). These differences are the result of environment, genetic factors or a combination of the two. Studies have showed a large environmental component, to most traits that distinguish between dwarf and large benthic charr in the lake. However it is unlikely that these traits are caused solely by the environment, meaning that some underlying genetic differences between the morphs are also present, and plays a role (4).

One type of relatively easily observed genetic differences between the morphs are single nucleotide polymorphism (SNPs) that can be assessed from RNA-sequencing data. Such data are available from three of the morphs from Lake Thingvallavatn (LB, SB and PL), taken from RNA-sequencing with Illumina technology of embryos at several stages of early development (Johannes Gudbrandsson-unpublished data).

Genetic variation

The genomes of the different morphs of Arctic Charr, are not identical reflecting both changes in individual bases (SNPs) or longer stretches of DNA. The RNA-transcriptome (Johannes Gudbrandsson-unpublished data) have previously identified candidate polymorphisms. Some of them had strong separation between morphs, meaning the allele frequency differed by more than 50 % between the samples. This means that a morph (for instance SB, PL or LB) can have a nearly fixed SNP in a gene, in which the other morph types do not. The SNPs are therefore different in certain genes of the different morphs. By also using estimates of gene expression (from the RNA transcriptome) it is also possible to ask about overlap of genetic and expression divergence. That may reflect that specific changes in a gene could influence variation in its expression, or a combination of *cis* and *trans* effects may shape the expression differences between morphs.

Aims

In this project I aimed to identify genes that associate with SB or benthic morphology in lake Thingvallavatn. I also wanted to test if the SNPs that associate with SB or benthic morphology in Lake Thingvallavatn also associate with SB in other populations. In order to do this, the SNPs first had to be identified searching for certain criteria (see Materials and Methods). Afterwards primers had to be made, to test the genes experimentally.

Testing which genes both have consisting genetic difference and expression difference among the morphs, and if there is a difference in number of genes fulfilling the criteria between the morphs, was pursued using restriction enzymes to cut the genes, and hereby genotype SNPs in differentially expressed genes in the morphs.

Materials and methods

Bioinformatic analyses of genetic and expression differences

From RNA-transcriptome data of three arctic charr morphs (SB, LB and PL) candidate SNPs had been estimated. My starting point was a list of 1428 SNP with the strongest genetic separation between morphs. I set out to summarize those SNPs by comparing allele frequencies in one morph vs. the other two. And by comparing two estimates of expression differences between the same pairs of morphs. These tables were made by running R-scripts, filtering the SNPs and genes for certain criteria. One set of criteria, were that the morphs should differ both genetically and at the expressional level, the p-value should be statistically significant, <0.5. Also there should be above 30% difference in the allele frequency of the SNPs. The data consisted of the three morphs LB, SB and PL which lead to 6 different tables. Two for each morph, one reporting on genes with morph effect on expression and one reporting genes with Morph by Time interaction in expression. Below is table M vs two interaction where the SB is compared with LB and PL (Murta) according to interaction. For the other 5 tables, see Appendix A.

DNA for population genetics

A total of 32 individuals, 8 x PL, 8 x SB, 8 x LB and 8 x PI, were analysed in the laboratory, see Table 1. The genes that we analysed were from the tables and some extra genes Arnar Pálsson found. The genes we tested were found by running the genes from the tables through a script written by Isak M. Johannesson, were available restriction enzymes cut in the area of a possible SNP. If there were no SNP, the enzymes would cut, and if there was a SNP, the enzymes would not. When starting the project, the DNA had already been isolated from the fish and stored in the freezer.

Table 1: 8 individuals from four different morphs used in this project.

Morph	**Individual number**							
SB	09	10	11	12	13	14	15	16
LB	01	02	03	04	05	06	07	08
PL	06	17	18	27	33	34	46	58
PI	09	10	11	12	13	14	15	16

Primers design

The primers (see appendix D) were made using Salmon contigs as template with the inserted fixed SNPs. Subsequently these were blasted against the normal salmon contigs using http://salmondb.cmm.uchile.cl/blast/, in order to avoid possible intron regions, which might contain the targeted sequence as we don't know their genetic code. The start and stop position for the future primers were chosen, and together with the position of SNP inserted on the website http://bioinfo.ut.ee/primer3/, which afterwards provided us with primers.

PCR

1 mastermix for each fragment was prepared. Each mastermix consisted of 13 μL diluted water, 2 μL of dNTPs, 2 μL buffer, 0.4 μL right primer, 0.4 μL left primer (primers, see Table 2) and 0.2 μL of taq polymerase. It all was mixed gently and hereafter 18 μL of Master Mix was mixed in an eppendorf tube with 2 μL sample of DNA.

The samples were incubated 95°C 5 min, in the PCR machine and ran 35 cycles of 45 seconds at 94°C, 51°C/53°C/55°C at 45 seconds and 1 min at 70°C. After the last cycle the samples were incubated furthermore for 10 min. at 72°C, and then stored at 12°C.

To check the quality of the PRC products before digestion with enzyme, the products were run on a 1.2% agarose gel. The gel was made of 0.6 g agarose heated with 50 mL 1xTAE buffer in a microwave, until the agarose was fully melted and the liquid was transparent. Afterwards 2.5 μL Ethidium Bromide was added. The mixture was cooled down, and poured into gel trays to set. The gel was loaded with 2 μL ladder, and 5 μL of the product mixed with 5 μL loading dye, subsequently run for 30 min at 100V.
The successful PCR products, with clear bands, were used further for digestion.

Restriction digestion of Arctic charr samples

The remnant of the PCR product (~15 μL) was digested with 2 μL of Enzyme and 2 μL of buffer. Then set to digest 5 hours at 37°C. The digestion products were immediately put on ice, and the loaded in a 2% gel, made by the exact same procedure as earlier, but with 1 g of agarose pr. 50 mL 1xTAE buffer, instead of 1.2% agarose. The gels were run at 80V for 1.5 hours, as lower voltage makes it more visible if the bands have separated.

The restriction enzymes were chosen after which we had in stock and would cut alleles with no SNP presented. The ladder was chosen after the sizes of the fragments, which made the 100bp ladder fit for these experiments.

Results

The overall objective of the study was to identify genes that associate with SB or benthic morphology in lake Thingvallavatn, using restriction enzymes to genotype SNPs in differentially expressed genes in the morphs.

The data summarized in six tables reflect correspondence of genetic and expression differences in the different morphs. The numbers of SNPs, varies between the categories, as is illustrated in Table 3.

Table 3: Numbers of SNPs presented in the six different tables.

Table	No. of SNPs
LB vs SB and PL (int)	259
LB vs two morph	393
PL vs two int	61
PL vs two morph	165
SB vs two int	87
SB vs two morph	204

As Table 3 illustrates, there is a clear variation in number of SNPs between the specific comparisons of morphs. Especially do comparisons involving LB have a lot more SNPs than the other two. This is because more genes are differently expressed between LB and either of the two other morphs, than are between SB and PL.

I also wanted to know how many SNPs and genes with strong separation between morphs, had no evidence of expression differences between any of the three morphs. By using bioinformatic filtering (p-value above 0.05 for each of the six expression comparisons) I found 514 SNPS, see appendix B. These genes have a genetic difference, but no expression differences in comparisons of those morphs.

I also wanted to look at how the 1428 SNPs distributed on genes, i.e. were there usually only one SNP per gene or were many SNPs per gene more common. Table 4 illustrates the distribution of SNPs in genes.

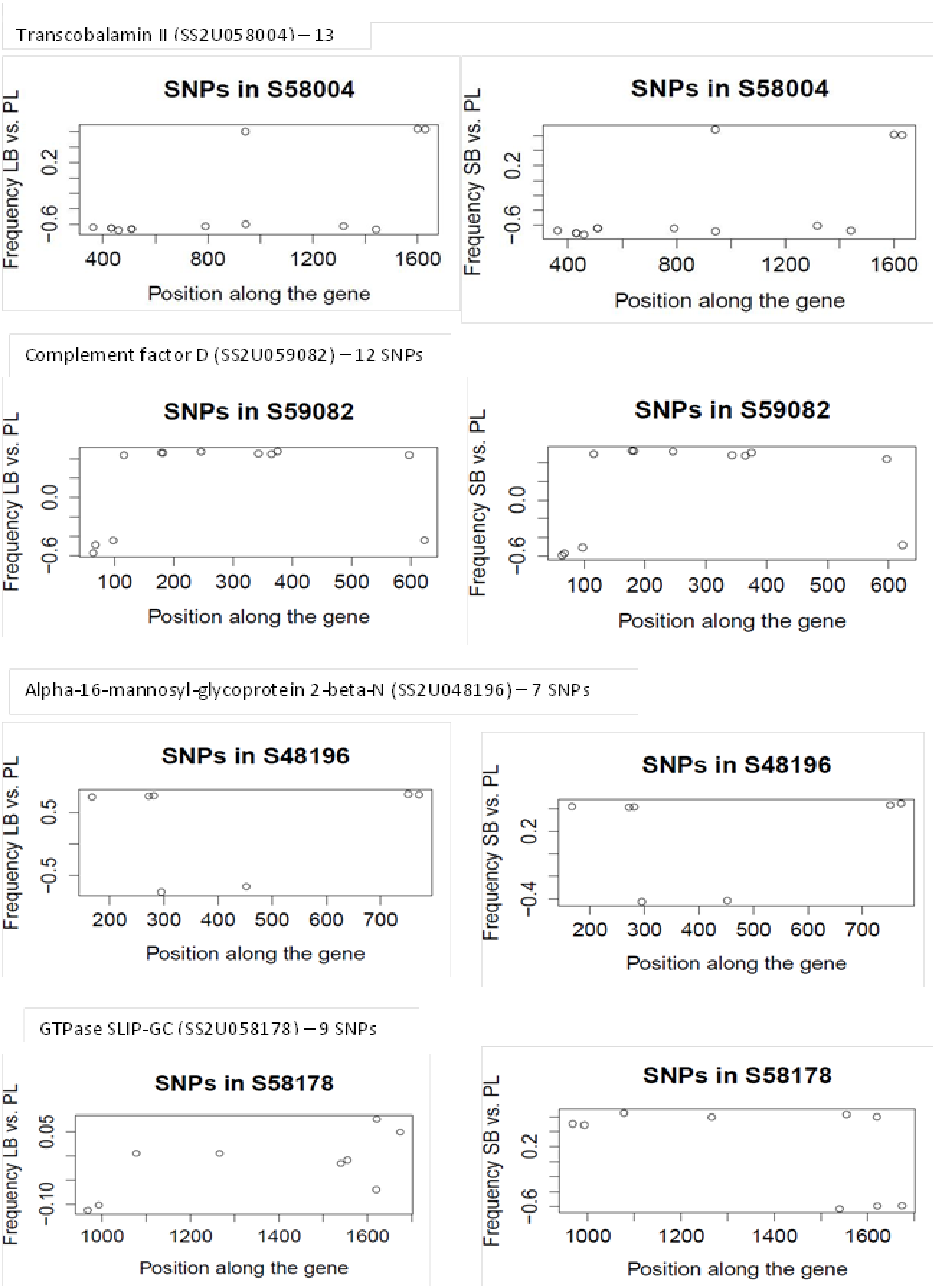
Table 4: Distribution of SNPs in genes.

No. of SNPS	1	2	3	4	5	6	7	9	10	12	13
No. of genes	739	170	38	18	10	5	5	1	1	1	1

Several genes have high number of SNPs, one gene contains 13 SNPs. I looked more carefully at the location of those SNPs within the contigs representing those genes, focusing on genes with 7 or more SNPs. This allows me both to look at the location of SNPs within the genes, and as the SNPs are relative to the salmon version of the gene, to look for indications of different rates of evolution on the SB, LB or PL branches. The data suggest that in some of the genes containing the highest numbers of SNPs may not be equally dispersed on the evolutionary branches separating the morphs. For instance at the (Transcobalamin-II) gene (SS2U058004) which has 13 SNPs, only three SNPs have derived alleles at high frequency in PL, while the other 10 are higher in SB and LB. Genes with such seemingly biased distribution of the SNPs in between the morphs can indicate that the respective gene has been through faster evolution in one morph or evolutionary branch. The transcobalamin II gene that bind and transport cobalamins including vitamin B12.

Gene SS2U059082 including 12 SNPs, 4 of the derived allele is higher in LB and SB and 8 is higher in PL, codes for a complement factor D. Gene SS2U048196 including 7 SNPs codes for Alpha-16-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase, and gene SS2U058178 including 9 SNPs, codes for GTPase SLIP-GC.

In some cases, the SNPs may cluster in certain parts of a gene or may not be present in some part of a gene, but as no formal statistical test was conducted then, I choose not to elaborate on that further.



Transcobalamin II (SS2U058004) – 13

Complement factor D (SS2U059082) – 12 SNPs

Alpha-16-mannosyl-glycoprotein 2-beta-N (SS2U048196) – 7 SNPs

GTPase SLIP-GC (SS2U058178) – 9 SNPs

Figure 2: Distribution of SNPs in genes with highest SNP number.

Table 5: SNPs with strong genetic and expression (interaction) difference between PL and the two benthic morphs.

Unigene.ID	Gene_name	Effect	Position	Ref	Var	Freq_LB	Freq_M	Freq_SB	Pad_SBvsM	
SS2U027681	Putative Peptide prediction	synonymous	448	G	T	0.0048	0.6587	0.6253	0.0482	0.0037
SS2U027681	Putative Peptide prediction	3prime	66	C	T	0.0191	0.6405	0.6175	0.0482	0.0037
SS2U027808	Collagen alpha-1(II) chain	synonymous	429	C	T	0.7522	0.2238	0.6306	0.0440	0.0000
SS2U027808	Collagen alpha-1(II) chain	synonymous	384	G	A	0.0342	0.6405	0.1674	0.0440	0.0000
SS2U029556	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	3prime	23	A	C	0.6573	0.9842	0.9610	0.0329	5,44E+07
SS2U031094	Apolipoprotein B-100	L to H / Hydrophobic to Positive	245	A	T	0.8872	0.2730	0.6729	0.0319	0.0243
SS2U033265	B8JI89 Novel protein similar to vertebrate collagen type VI alpha 3 (COL6A3) (Fragment)	A to P / Hydrophobic to Special	627	G	C	0.1278	0.5652	0.0037	0.0357	0.0088
SS2U033465	Pancreatic secretory granule membrane major glycoprotein GP2	3prime	467	T	A	0.8366	0.3066	0.3913	0.0281	0.0228
SS2U037160	Tricarboxylate transport protein mitochondrial	5prime	67	G	T	0.0000	0.0292	0.1886	0.0281	0.0368
SS2U041872	Filensin	3prime	431	G	A	0.2757	0.8281	0.2884	0.0210	5,13E-03
SS2U041872	Filensin	3prime	479	A	G	0.7084	0.1652	0.6228	0.0210	5,13E-03
SS2U041872	Filensin	3prime	261	T	A	0.2661	0.9283	0.2941	0.0210	5,13E-03
SS2U042555	Putative Peptide prediction	3prime	482	G	T	0.8633	0.3542	0.8523	0.0492	0.0308
SS2U048214	Translation initiation factor IF-2	S to P / Polar to Special	1199	A	G	0.0543	0.6272	0.3163	0.0282	7,06E+07
SS2U048214	Translation initiation factor IF-2	S to F / Polar to Hydrophobic	958	G	A	0.9357	0.2673	0.6996	0.0282	7,06E+07
SS2U048341	Sideroflexin-3	K to E / Positive to Negative	899	A	G	0.4421	0.4167	0.3595	0.0499	0.0250
SS2U048664	Papilin	3prime	651	A	T	0.8684	0.4399	0.9783	0.0013	0.0045
SS2U048664	Papilin	A to S / Hydrophobic to Polar	333	G	T	0.8920	0.5100	0.9907	0.0013	0.0045
SS2U048664	Papilin	A to P / Hydrophobic to Special	117	G	C	0.1098	0.5716	0.0125	0.0013	0.0045
SS2U048801	Putative Peptide prediction	5prime	301	C	T	0.3400	0.9314	0.5360	0.0011	0.0092
SS2U049381	UPI0000F211FE PREDICTED: solute carrier family 39 (zinc transporter) member 5	3prime	720	A	G	0.2845	0.2266	0.7686	0.0342	2,12E+09
SS2U049425	Tropomyosin alpha-4 chain	synonymous	716	C	T	0.9984	0.6354	0.8984	0.0085	0.0158
SS2U049425	Tropomyosin alpha-4 chain	G to S / Special to Polar	489	G	A	1.0000	0.6810	0.9197	0.0085	0.0158
SS2U049839	Complement C3	Q to H / Polar to Positive	758	G	T	0.0006	0.0002	0.2517	0.0038	3,87E+07
SS2U050200	Pogo transposable element with ZNF domain	K to M / Positive to Hydrophobic	623	A	T	0.9045	0.9833	0.2685	0.0065	0.0338
SS2U050200	Pogo transposable element with ZNF domain	synonymous	864	C	T	0.0606	0.0057	0.3195	0.0065	0.0338
SS2U050646	UPI00017B55A8 UPI00017B55A8 related cluster	3prime	730	T	G	0.0017	0.0020	0.2409	0.0013	3,79E+09
SS2U050646	UPI00017B55A8 UPI00017B55A8 related cluster	3prime	777	A	G	0.1147	0.6639	0.8990	0.0013	3,79E+09
SS2U051189	Putative Peptide prediction	synonymous	454	G	A	0.0543	0.0898	0.4190	0.0003	0.0002

SS2U051545	Glycerophosphodiester phosphodiesterase 1	synonymous	647	C	T	0.0000	0.2912	0.0291	0.0063	3,29E+09
SS2U051564	Proteolipid protein 2	A to V / Hydrophobic to Hydrophobic	308	C	T	0.0143	0.5488	0.0888	0.0123	0.0156
SS2U051564	Proteolipid protein 2	L to M / Hydrophobic to Hydrophobic	544	C	A	0.0128	0.5088	0.0436	0.0123	0.0156
SS2U052022	Protein TsetseEP	5prime	977	T	C	0.2590	0.1567	0.0651	4,74E+09	2,20E+07
SS2U052089	Putative Peptide prediction	5prime	310	C	G	0.9889	0.5279	0.9627	0.0233	0.0005
SS2U052089	Putative Peptide prediction	C to Y / Special to Hydrophobic	672	G	A	0.0024	0.4929	0.0186	0.0233	0.0005
SS2U052089	Putative Peptide prediction	3prime	919	T	C	0.0257	0.7964	0.0698	0.0233	0.0005
SS2U052089	Putative Peptide prediction	5prime	339	G	A	0.9825	0.5053	0.9617	0.0233	0.0005
SS2U052425	AT-rich interactive domain-containing protein 4A	3prime	464	C	A	0.3671	0.0250	0.5921	0.0091	0.0000
SS2U053194	DNA polymerase zeta catalytic subunit	3prime	1677	T	C	0.2718	0.2895	0.2148	0.0283	5,93E+03
SS2U053194	DNA polymerase zeta catalytic subunit	synonymous	962	G	A	0.1642	0.6878	0.5385	0.0283	5,93E+03
SS2U053510	MKL/myocardin-like protein 1	3prime	144	C	A	0.8420	0.2927	0.4059	0.0263	0.0023
SS2U053514	UPI00016E0D66 UPI00016E0D66 related cluster	3prime	343	G	A	0.8238	0.1998	0.3080	0.0012	0.0004
SS2U053514	UPI00016E0D66 UPI00016E0D66 related cluster	synonymous	167	G	A	0.8930	0.3002	0.5198	0.0012	0.0004
SS2U053732	Probable E3 ubiquitin-protein ligase TRIP12	3prime	684	G	A	1.0000	0.9738	1.0000	0.0009	6,48E+04
SS2U053784	Creatine kinase M-type	T to A / Polar to Hydrophobic	2245	T	C	0.1637	0.5241	0.6524	0.0017	0.0249
SS2U054187	Lysine-specific demethylase 5B	K to R / Positive to Positive	1310	T	C	0.0094	0.1329	0.5450	0.0050	0.0004
SS2U054187	Lysine-specific demethylase 5B	3prime	359	A	C	0.0267	0.0802	0.5428	0.0050	0.0004
SS2U054187	Lysine-specific demethylase 5B	A to S / Hydrophobic to Polar	1425	C	A	0.9663	0.8950	0.3014	0.0050	0.0004
SS2U054344	Selenoprotein Pb	3prime	1514	G	A	0.0254	0.8130	0.2818	0.0160	2,70E+09
SS2U054344	Selenoprotein Pb	3prime	1494	G	A	0.9723	0.1809	0.7532	0.0160	2,70E+09
SS2U055468	Metalloproteinase inhibitor 2	K to R / Positive to Positive	788	T	C	0.9430	0.6215	0.2733	0.0077	1,06E-04
SS2U055653	Putative Peptide prediction	5prime	1336	T	A	0.2170	0.0078	0.4297	0.0017	0.0242
SS2U056095	Eukaryotic translation initiation factor 4 gamma 1	synonymous	1303	C	T	0.0133	0.4502	0.3457	0.0087	4,15E+08
SS2U056124	Globoside alpha-13-N-acetylgalactosaminyltransferase 1	R to C / Positive to Special	837	G	A	0.3210	0.2403	0.8231	6,08E+09	9,81E+09
SS2U056450	Myomesin-1	3prime	1186	G	C	0.3424	0.9018	0.3326	0.0047	0.0433
SS2U056641	Beta-taxilin	3prime	1338	T	G	0.9965	0.5300	0.9985	0.0398	0.0319
SS2U057930	Glucose-6-phosphate isomerase	3prime	811	G	T	0.2881	0.1000	0.6935	0.0001	0.0007
SS2U058206	Kunitz-type protease inhibitor 2	synonymous	1330	C	T	1.0000	0.5945	0.9546	0.0172	3,00E-09
SS2U058694	Intermediate filament protein ON3	3prime	367	T	A	0.1702	0.1159	0.5794	0.0257	0.0142
SS2U059309	Selenoprotein Pa	3prime	1531	T	A	0.0017	0.0498	0.4811	0.0105	0.0276
SS2U059309	Selenoprotein Pa	3prime	2097	T	A	1.0000	0.5720	0.9574	0.0105	0.0276

PCR genotyping of SNPs in Arctic charr genes

To find the right temparatures for the primers, 2 degrees were subtracted from the lowest melting temperature of each primerpair. The genes and belonging primers were hereafter organized in groups according to the temperature when running the PCR. Table 2 represent the PCR groups with their PCR running temperature, and also the quality of the PCR product. Information on the location of the polymorphisms in the genes that we aimed to genotype are detailed below, see Table 6.

Table 2: Primers and PCR product quality. The ones that were digested are marked with an asterix.

Contig	**Sequence**	**Primer**	**Tm**	**TA**	**PRC group**	**Band quality**
SS2U003354*	ACAGTGTAGAGCCAGTCGTC	F1	57.5	55	3	Good
SS2U003354*	AGCAGGTCTAACTCATCCAGG	R1	57.4	-	-	-
SS2U036171	TTGAATGTGACACCAGCACG	F1	55.4	53	2	Poor
SS2U036171	GTGCTCTGGGACTTTGTGTT	R1	55.4	-	-	-
SS2U041872*	AGAACTGAGTCTGCGATAAGGT	F1	55.6	53	2	Good
SS2U041872*	TGGAACCAGCTAACACTTGC	R1	55.4	-	-	-
SS2U044679*	TCCTCCTGTCCTAATGGCAC	F1	57.5	53	2	Good
SS2U044679*	TGAATCAACATGCCAGGCTG	R1	55.4	-	-	-
SS2U044964*	GTCTACCCACTTTCCCACCA	F1	57.5	53	2	Good
SS2U044964*	TGACATTGTGTGGGGTGGTA	R1	55.4	-	-	-
SS2U047101*	GTTCCTATCCCCAGCAAGCT	F1	57.5	55	3	Good
SS2U047101*	GCGTCTGAATAAGGTGGTGC	R1	57.5	-	-	-
SS2U049150	ACAAATGAGGTGAGATGGCA	F1	53.4	51	1	Poor
SS2U049150	TGGTCCAAGGCAGTCAAAGA	R1	55.4	-	-	-
SS2U055020	GACCCACTATGAGTCGTCCT	F1	57.5	55	3	Poor
SS2U055020	GCTCTCCTCTTTTCTGTGGC	R1	57.5	-	-	-

SS2U056748	*AATGCCTTGTGAACACCAGC*	*F1*	*55.4*	*53*	*2*	*Poor*
SS2U056748	*GAGTCGGTCTTCCAAACACT*	*R1*	*55.4*	*-*	*-*	*-*
*SS2U056815**	*GACCCTATACTGACCCGCAT*	*F1*	*57.5*	*51*	*1*	*Good*
*SS2U056815**	*AACAACGCTGCAGTCAAATG*	*R1*	*53.4*	*-*	*-*	*-*
*SS2U057528**	*CTCCAGGTGTTGTGCTGAAC*	*F1*	*57.5*	*53*	*2*	*Good*
*SS2U057528**	*CGCCCCTGTGTTTTGAAGAT*	*R1*	*55.4*	*-*	*-*	*-*
*SS2U059004**	*TGTTCAAAGGCATGGCAAATTG*	*F1*	*54*	*51*	*1*	*Fine*
*SS2U059004**	*GCGTGTATGTTTTACTTTGCTGT*	*R1*	*54.1*	*-*	*-*	*-*

As seen in Figure 3 the PCR products were diffused. This is possibly because other regions, in addition to or including the target region, are being amplified. This makes the product weaker, and can therefore be difficult to see after digestion. Unfortunately this was the case for several of the PCR products, see Appendix C.

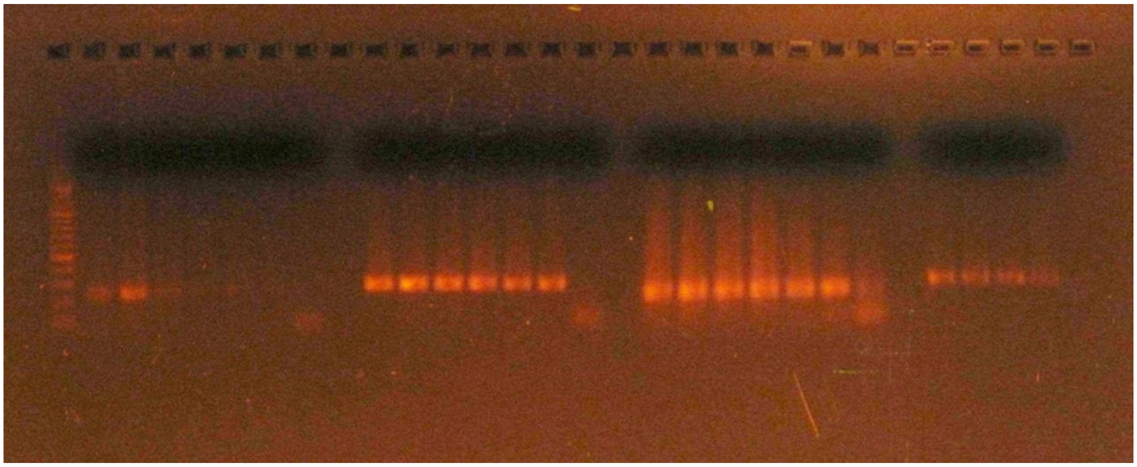


Figure 3: Diffused PCR products of individual PL34, PL58, SB13, SB15, LB05, and LB07 from fragment, from the right; SS2U036171, SS2U041872, SS2U044679 and SS2U044964   group 2.

The cause of this is most likely that the temperature in the annealing step is too low in order for the hybridization of the primer to be specific, hence the diffused bands.

Therefore I chose to test the optimum temperature for the primers belonging to each gene/SNP on DNA from one individual. The individual was D09 (see Figure 4) and the gene tested was SS2U044964, as this gene and individual had showed results both in PCR and digestion reactions.

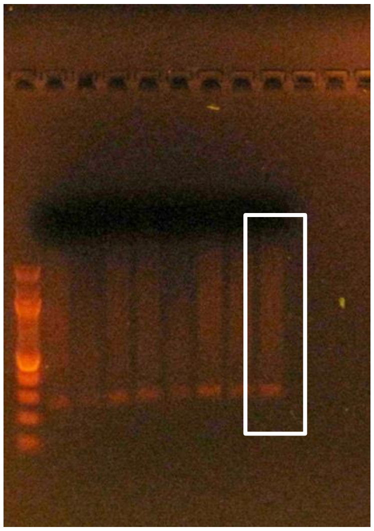


Figure 4: Eight individuals of the SB morph at temperature 53°C. Individual D09, marked in white, was used for further testing.

The procedure was run exactly as before, but with the DNA mixed in the Mastermix tube, to be completely sure the same concentration of DNA was present in each well when running the PCR. The PCR program was run as before, but instead with an annealing temperature of 51°C, 53°C, 55°C, 57.5°C, 60°C and 62.5°C, see Figure 5.

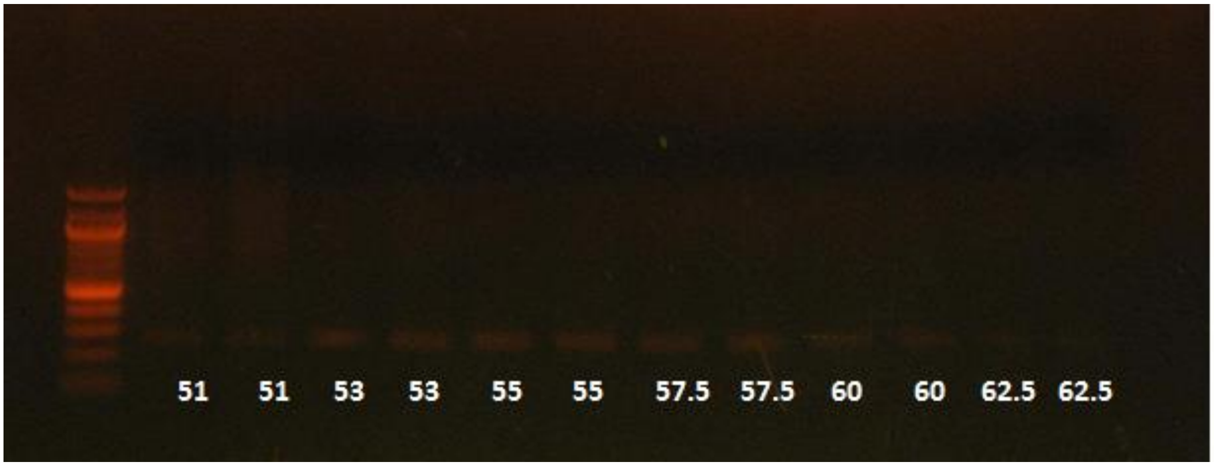


Figure 5: Temperature test of the primers for SNP located in SS2U044964. The primers were tested at following temperatures; 51°C, 53°C, 55°C, 57.5°C, 60°C and 62.5°C.

Figure 5 shows clearly dilution of the bands at temperature 51°C. The bands at temperature 53°C, 55°C and 57.5°C are very similar. When it is showed in Figure 4 that with a temperature at 53°C, the bands are diffused, it would be better to use a higher temperature than 53°C.

However, PCR products for SS2U003354, SS2U041872, SS2U044679, SS2U044964, SS2U047101, SS2U056815, SS2U057528 and SS2U059004 were deemed good enough, to digest the product. The genes that had good enough PCR products to run a digestion, are marked by asterix in Table 2.

Restriction enzyme digestions to genotype SNPs

The estimated length of the amplified fragments for the digested and undigested alleles are shown in Table 6.

Table 6: Gene lengths before and after cutting with restriction enzymes.

Contig	SNP location	Start	Fragment length undigested	Fragment length digested
SS2U003354*	253	178	132	75
SS2U003354*	253	310	132	57
SS2U036171	242	113	228	129
SS2U036171	242	341	228	99
SS2U041872*	261	156	242	105
SS2U041872*	261	398	242	137
SS2U044679*	481	337	220	144
SS2U044679*	481	557	220	76
SS2U044964*	613	475	245	138
SS2U044964*	613	720	245	107
SS2U047101*	202	156	99	46
SS2U047101*	202	255	99	53
SS2U049150	687	562	157	125
SS2U049150	687	719	157	32
SS2U055020	1092	943	312	149
SS2U055020	1092	1255	312	163
SS2U056748	1293	1228	195	65
SS2U056748	1293	1423	195	130
SS2U056815*	1294	1111	374	183
SS2U056815*	1294	1485	374	191
SS2U057528*	851	752	153	99
SS2U057528*	851	905	153	54
SS2U059004*	1612	1395	270	217
SS2U059004*	1612	1665	270	53

The first PCR product I digested was amplified SS2U044964, using the *SacI* enzyme (Figure 6).

Figure 6 shows the undigested fragment of SS2U044964 is 245bp, while the digested part, yields two bands that are estimated to be 138 and 107bp. The results of the digestion on two individuals from SB, LB and PL respectively, are in agreement with Figure 6.

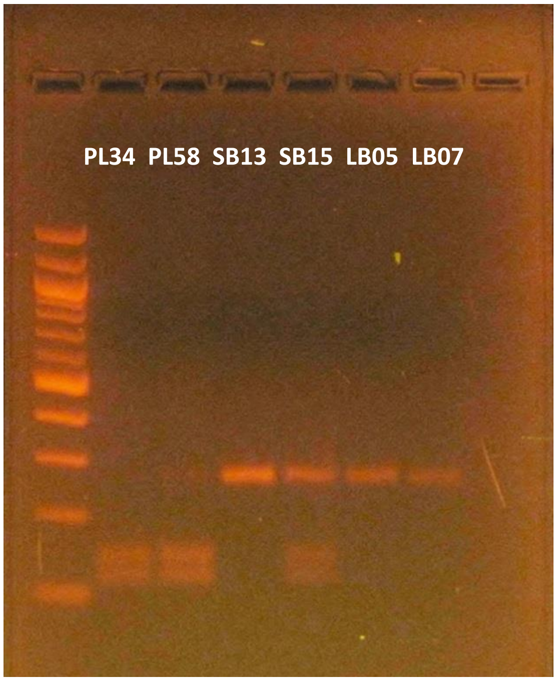


Figure 6: Digestion with SacI of a fragment of SS2U044964.

After the good results of digesting the fragment of SS2U044964, this fragment was amplified from eight different individuals of the respective four morphs SB, LB, PL and PI (32 individuals in total) and digested, this can be seen in Figure 7 and Figure 8.

The digestion of fragments from SB and LB, see Figure 7, did not gave very clear results, as to why it in some cases can be difficult to determine whether the candidate SNPs can be verified. There are four SB individuals where both the alleles were cut in respect to the gene SS2U044964 including SB 11, SB 12, SB 13 and SB 14. Three of the individuals, SB 10, SB 15 and SB 16, are heterozygous, with one allele digested, while SB 09 didn't have any allele digested.
Also four of the LB individuals, LB 02, LB 05, LB 06, LB 07, were homozygous with digested alleles

with respect to the gene, while three individuals, LB 03, LB 04 and LB 08 were heterozygous. LB 01 was not possible to genotype from these results.

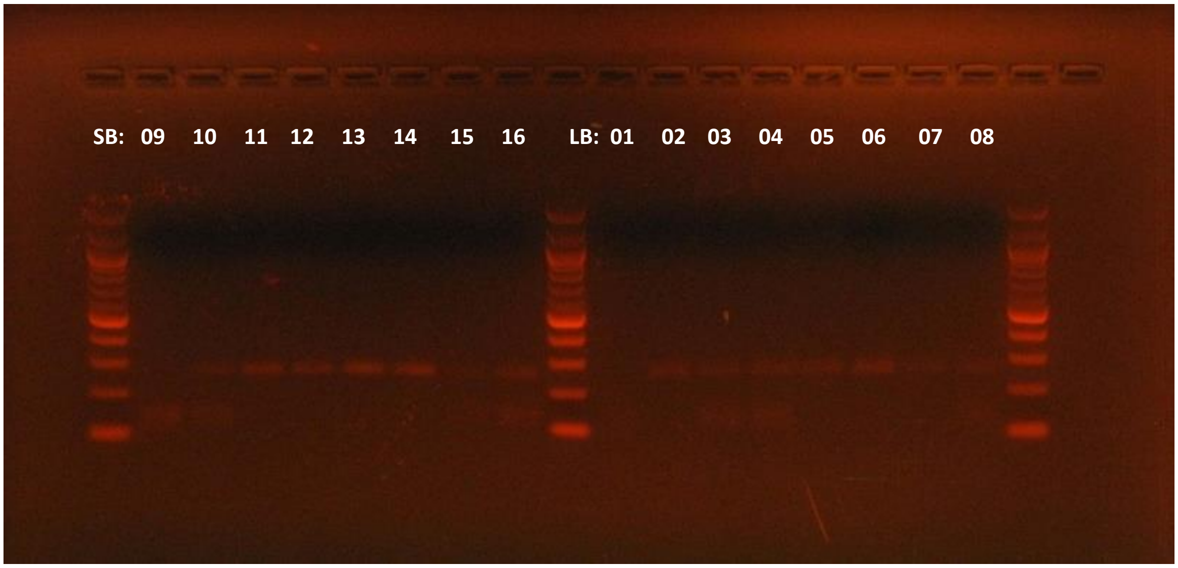


Figure 7: Digestion of SS2U044964 respectively eight SB and LB individuals with SacI.

In Figure 8, we see that none of the PL or PI individuals are homozygous with digested alleles. Four of the PL individuals, PL 06, PL 17, PL 27 and PL 46 are heterozygous, while the last four, PL 18, PL 33, PL 34 and PL 58 only had undigested alleles.

One PI individual, PI 16 is heterozygote with one allele digested in this gene, whereas the last seven, PI 09, PI 10, PI 11, PI 12, PI 13, PI 14 and PI 15 had both alleles undigested.

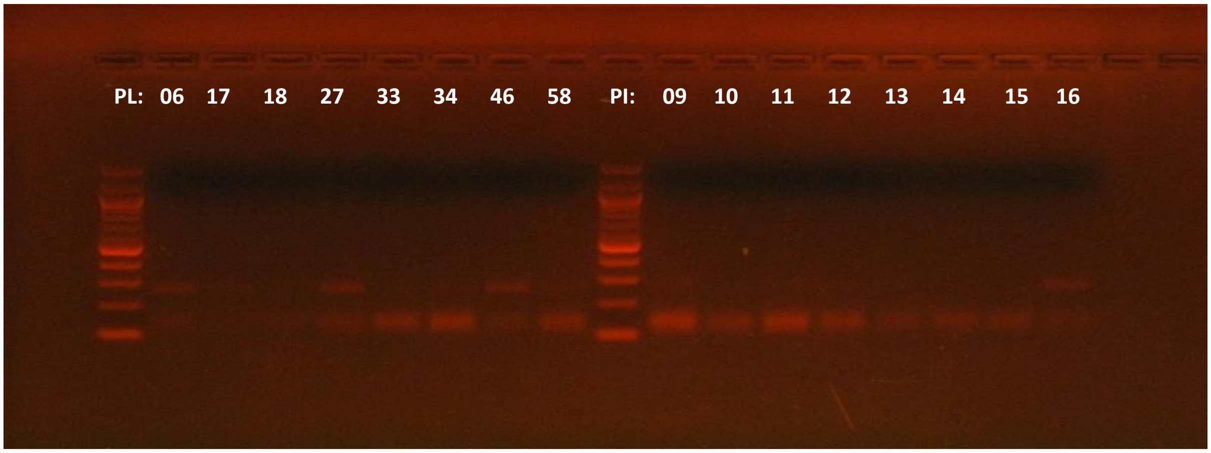


Figure 8: Digestion of respectively eight PL and PI individuals with SacI.

Also the SNP at bp 1612 in gene SS2U059004 gave fine PCR results, and was digested with the restriction enzyme EcoRI, see Figure 9.

The undigested fragment of SS2U059004 is 270bp, while the digested part, yields two bands that are 217 and 53bp. The results of the digestion on two individuals from SB, LB and PL respectively, are in agreement with Figure 9.

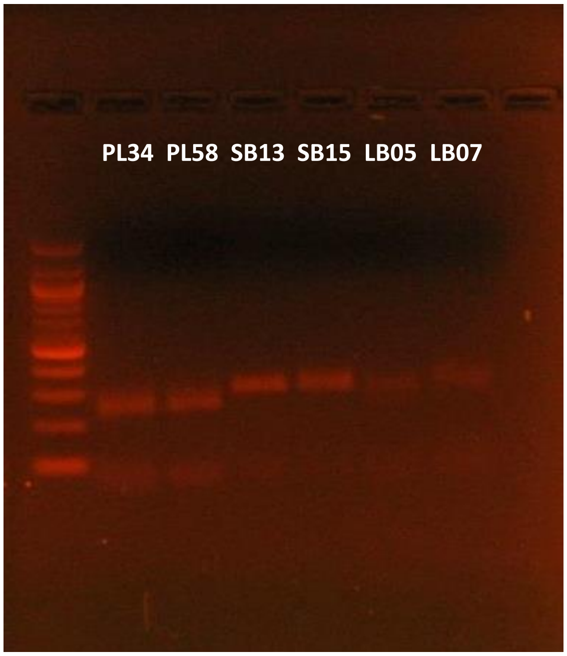


Figure 9: Digestion of 6 individuals with EcoRI.

If the restriction enzyme cut the gene, one of the fragments will only be 53bp long. Those small fragments have probably run out of the gel, as it is not to be seen on the gels, even though some of the genes have been cut, see Figure 10 and Figure 11.

Figure 10 shows six individuals, SB 10, SB 11, SB 12, SB 13, SB 14 and SB 15 that are homozygous with both alleles digested in gene SS2U059004, and two individuals, SB 09 and SB 16, that are heterozygous.

Three of the LB individuals, LB 03, LB 04 and LB 06 are homozygous with both alleles digested, while one is heterozygote, and one is homozygote without no alleles digested. The genotype of the remaining three cannot be determined from this experiment.

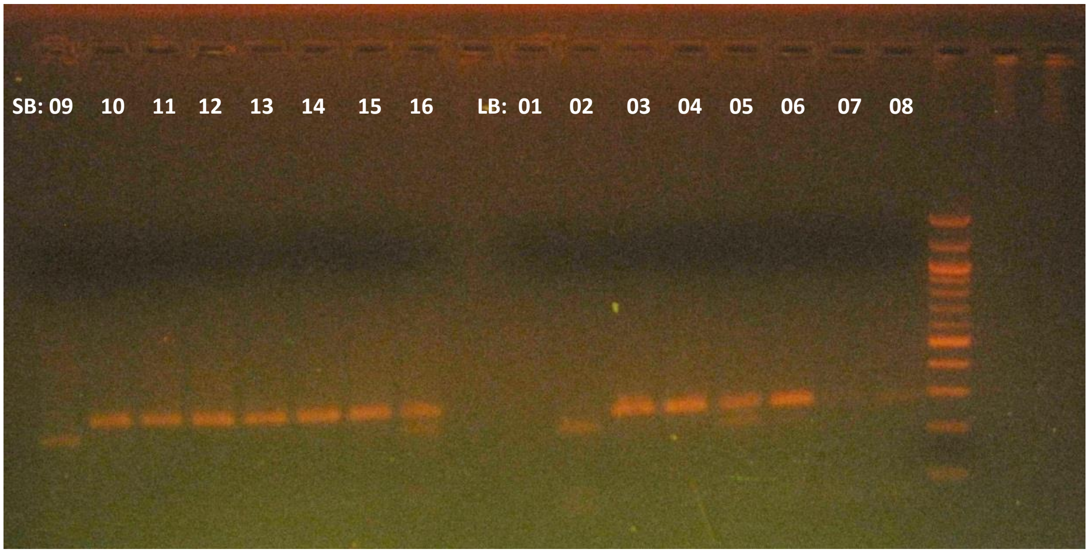


Figure 10: Digestion of respectively eight SB and LB individuals with EcoRI SS2U059004.

Figure 11 shows the EcoRI digestion of fragments from PL and PI in gene SS2U059004.

One PL individual, PL 33 had both alleles digested, three individuals; PL 17, PL 18 and PL 58 were heterozygous while four of the individuals, PL 06, PL 27, PL 34 and PL 46 were heterozygous with no alleles digested see Figure 11.

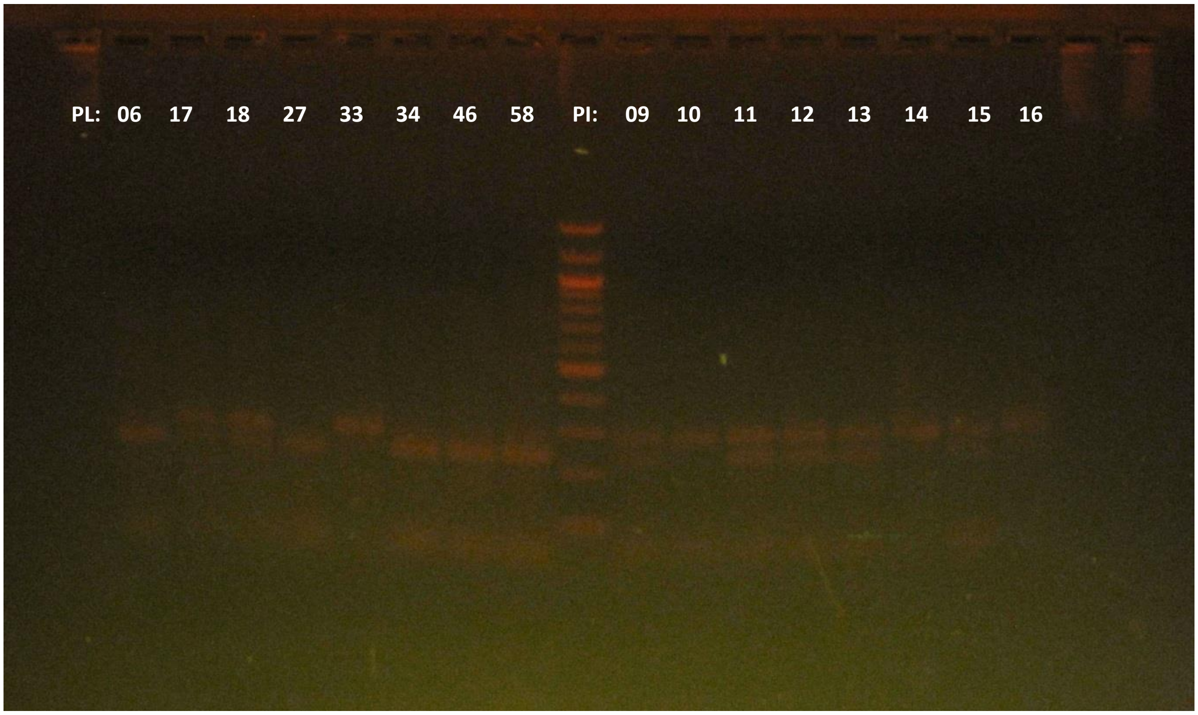


Figure 11: Digestion of respectively eight PL and PI individuals with EcoRI SS2U059004.

The raw data represented in Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11 is summarized in Table 7 below.

The data demonstrate that these two variations in genes SS2U044964 and SS2U059004 that were found in the transcriptome are real. Secondly the data show that both variants segregate with the craniofacial morphology of the fishes. Both variants are more similar among the benthic (SB and LB) on one hand, and the limnetic (PI and PL) fishes on the other hand.

I was also interested in knowing how similar the estimated allele frequencies were in the transcriptome and the population samples. The data show very good agreement between the two datasets, suggesting the transcriptome of pooled individuals may be useful for finding real genetic signals separating the morphs.

In Table 7 and Table 8 all the individuals and their number of alleles containing the SNP are illustrated.

Table 7: Numbers of alleles containing the 2 SNPs, in SS2U044964.

PI	Alleles 4964	SB	Alleles 4964	LB	Alleles 4964	PL	Alleles 4964
09	0	09	0	01	NA	06	1
10	0	10	1	02	2	17	1
11	0	11	2	03	1	18	0
12	0	12	2	04	1	27	1
13	0	13	2	05	2	33	0
14	0	14	2	06	2	34	0
15	0	15	1	07	2	46	1
16	1	16	1	08	1	58	0
Calculated frequency of SNP			0.688		0.79		0.25
Table frequency of SNP			0.730		0.633		0.075

NA: Data not available

Table 8: Numbers of alleles containing the 2 SNPs, in SS2U059004.

PI	Alleles 9004	SB	Alleles 9004	LB	Alleles 9004	PL	Alleles 9004
09	1	09	1	01	NA	06	0
10	1	10	2	02	0	17	1
11	1	11	2	03	2	18	1
12	1	12	2	04	2	27	0
13	1	13	2	05	1	33	2
14	2	14	2	06	2	34	0
15	1	15	2	07	NA	46	0
16	2	16	1	08	NA	58	1
Calculated Frequency of SNP			0.875		0.7		0.313
Table frequency of SNP			0.902		0.936		0.343

NA: Data not available

The calculated frequencies of SNPs in the genes are calculated from the numbers of alleles. These allele frequencies approximately correspond to the frequencies from the transcriptome. The allele frequencies of those other variants in the transcriptome can be found in Appendix A.

Discussion

The four morphs of Arctic Charr in Lake Thingvallavatn have evolved from the same ancestor in about 10,000 years (1). This makes the Arctic Charr from Lake Thingvallavatn quite unique, as the four morphs differ greatly in shape, size and ecology (1, 6). There are many lakes and river systems in Iceland, some quite a lot younger, where the Arctic Charr fish lives, but this many morphs are only found in Lake Thingvallavatn.

However, similar morphs are found in some lakes and ponds. In particular small benthic charr are found in many locations (6), showed that they had evolved independently. It is interesting if the Charr in these lakes also evolve to the exact same morphs, and if they develop the same way as in Lake Thingvallavatn, or through develop slightly differently.

By comparing the expression of genes and frequencies of SNPs in different morphs it might be possible to identify the genes and developmental systems that influence the phenotypic differences between the morphs. It is even possible that some of the variants relate to differences in the expression of the genes they are found in, but it is more likely that coupling of expression differences and genetic differences is due to other genetic changes in the regulatory regions of the relevant genes. Those cannot be found by looking at SNPs in the transcribed portions of the genome.

I attempted, in collaboration with Isak M. Johannesson to verify SNPs in 12 genes with PCR and digestion methods.

PCR amplification of the 12 selected genes from the different morphs of Arctic Charr was executed, and the products analysed with gel electrophoresis. Only eight genes gave a visible PRC product, and were continued in use for digestion. Unfortunately the temperature of the primers was only tested in the end of the project, as the temperature otherwise was set after the protocol following the primers. In general we had problems with bad PCR products, and gels. The reason for the bad PCR products may have been that the temperature in the annealing process was too low, which resulted in unspecific primers, which amplified unwanted fragments.

A problem could also be that the PCR primers did not bind to the Arctic Charr genome, due to other genetic changes.

Restriction enzyme digestion of the 8 different gene products only five genes showed results, among these only three were applicable. Hence, three of the fragments were successfully digested, while 2 of the fragments showed that a wrong PRC product had been amplified. The reason why only 3 out of 12 gene fragments were successfully digested can be the weak PCR products because of the wrong temperature. In 2 of the 3 failed digestions, it turned out that we had used the enzyme (*DpnI*), which only digests methylated DNA. Although we only had few, not always clear, results, the frequency of alleles in SS2U059004 and SS2U044964 genes from the three different morphs, corresponded to the allele frequencies estimated from the transcriptome. It would be interesting to see whether this would be valid for all the genes we studied.

Conclusion

By experimental and bioinformatic study of the four morphs in Lake Thingvallavatn, I have started to investigate the genetics of these morph and the nucleotide changes that may relate to their evolution and ecology. By looking at the SNPs that separate the morphs, and the expression of the genes that those SNPs reside in I see that some changes may be associated with certain morph types. Through this project I did not find out what that means for the exact phenotypes or ecological differences of those fishes. It is possible that some of these SNPs in the genes influence a certain trait in the morph. Through the experimental work, I can conclude that the frequencies in the transcriptome are generally consistent with the frequencies gained through experimental work. That indicates that this transcriptome gives valuable insight into the frequencies of SNP in the transcribed part of the Arctic charr genome. This is good for future studies of the genetics of the sympatric Arctic charr morphs from Lake Thingvallavatn.

Acknowledgements

I would like to thank my supervisor Arnar Pálsson for his guidance, help with data analysis and cheerful spirit. I also thank Kalina H. Kapralova for help in the laboratory and Isak M. Johannesson for good collaboration.

I furthermore appreciate the resources and facilities, University of Iceland and Department of Natural Sciences have put at my disposal for this project.

References

1. Jonsson B, Jonsson N. Polymorphism and speciation in arctic charr. Journal of Fish Biology, J.Fish Biol. 2001;58(3):605-38.

2. Ahi EP, Kapralova KH, Palsson A, Maier VH, Gudbrandsson J, Snorrason SS, et al. Transcriptional dynamics of a conserved gene expression network associated with craniofacial divergence in arctic charr. EVODEVO. 2014;5(1).

3. Kapralova KH, Franzdóttir SR, Jónsson H. Patterns of MiRNA expression in arctic charr development. PLoS ONE. 2014;9(8).

4. Sandlund OT, Gunnarsson K, Jonasson PM, Jonsson B, Lindem T, Magnusson KP, et al. The arctic charr salvelinus-alpinus in thingvallavatn. Oikos. 1992;64(1-2):305-51.

5. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson H, Grimwood J, et al. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science, Science. 2005;307(5717):1928-33.

6. Kapralova KH, Morrissey MB, Kristjánsson BK, Olafdóttir GÁ, Snorrason SS, Ferguson MM, Evolution of adaptive diversity and genetic connectivity in Arctic charr (Salvelinus alpinus) in Iceland. 2011;106(3):472-87