NVD Rit

SERRITGERÐ
Thesis

TØKNIFRÁGREIÐ
Technical Repo

UNDIRVÍSINGARTILFAR
Teaching Material

UPPRIT
Notes

BSc Thesis

Steps towards next generation sequencing of Haemoglobin

Ari Jespersen

NVD*Rit* 2014:05

NÁTTÚRUVÍSINDADEILDIN FRÓÐSKAPARSETUR FØROYA
Faculty of Science and Technology University of the Faroe Islands

Heiti / Title ***Steps towards next generation sequencing of Haemoglobin***

Menning av næstu generatión sekvensering av hemoglobin

Høvundar / Authors **Ari Jespersen**

Vegleiðari / Supervisor Svein-Ole Mikalsen, Fróðskaparsetur Føroya

Eyðfinn Magnussen, Fróðskaparsetur Føroya

Ábyrgdarvegleiðari / Responsible Supervisor Svein-Ole Mikalsen, Fróðskaparsetur Føroya

Ritslag / Report Type BSc ritgerð, lívfrøði

BSc Thesis, Biology

Latið inn / Submitted 5. september 2014

NVD*Rit* 2014:05

ISSN 1601-9741

Útgevari / Publisher Náttúruvísindadeildin, Fróðskaparsetur Føroya

Bústaður / Address Nóatún 3, FO 100 Tórshavn, Føroyar (Faroe Islands)

✆ • 🖷 • @ +298 352 550 • +298 352 551 • nvd@setur.fo

Contents

1. Introduction

1.1 Population genetics

Population genetics is the study of allele frequency distribution throughout a population, and how the distribution of the alleles can vary within and between subpopulations of a population (Hartl and Clark, 2007). Should the changes of DNA sequence result in any beneficial or detrimental effects, it will lead to a change in selection pressure on that gene. The selection pressure on a specific gene can vary between populations, depending on what environmental challenges the different populations face.

Population genetics is becoming an increasingly important for management of biological resources, including fish stock, because to be able to say with any confidence how a population is thriving, you need to be able to look at its subpopulations, and you can then set optimal strategies for management of said stock (Coyle, 1997).

Population genetics is also an important tool in other aspects, because it with genetic mapping can detect genes for known disease susceptibility, both in humans and other organisms (Hartl and Clark, 2007). This is hard to use on a wild population of fish because of migrations and the many subpopulations, but it is important in aquacultures.

1.2 Methods in population genetics

Mendel's experiments with plant hybridization is often said to be the foundation of genetics. However his experiments only focused on a visually observable difference in the phenotype. Phenotypic markers have been very important in classical population genetics, but in the last 50 years or so, a number of molecular methods have been introduced.

Among the first molecular methods, was the study of isoenzymes. Isoenzymes, which are enzymes that catalyze the same reaction, but have slightly different sequence of amino acids, can be used as a simple molecular marker through gel electrophoresis. Other proteins also have isoforms, among them haemoglobin. Haemoglobin was among the proteins used in early population genetic studies of cod, and some phenotypic alleles were discovered with this method (Sick, 1961).

A genetic marker is a known sequence of DNA that can be used to identify individuals or species. Genetic markers are commonly used in forensic science and paternity tests. Methods directly studying DNA are much more powerful than studies of proteins,

because with these methods we can see differences on the DNA sequence, and not just the differences that would result in a change of amino acids in the protein sequence (as for isoforms). The markers can be single nucleotide polymorphisms (SNPs), minisatellites or microsatellites. SNPs are single nucleotide variations on the DNA strand, and if within the borders of a gene, they create different alleles of the same gene. Microsatellites, also known as simple sequence repeats (SSR), are used as molecular markers. Microsatellites are repeating sequences of 2-6 base pairs and they can be repeated in tandem many times at the particular site (Hartl and Jones, 2006).

Both microsatellites and SNPs are great tools for population studies. In the foreseeable future, whole genome sequencing can give even more detailed resolution.

1.3 Cod

1.3.1 Atlantic Cod

Atlantic Cod (*Gadus morhua*) is found on the continental shelves and banks on both sides of the North Atlantic (Figure 1.1). Cod prefers depths of 150 – 200 m, but can be found below 600 m, its presence generally relies more on the amount of prey than on temperature (Magnussen, 1993).

Cod is important as a human food resource and is therefore also an economically important species for nations around the North Atlantic. The most spectacular collapse in the cod populations occurred around Newfoundland, where people had lived of cod fisheries for about 500 years, and in the 1980's the population was fished to near extinction through technological advances and lack of population management (Shelton, 2004). Even now 30 years later, studies show that the cod stock around Newfoundland, while recovering, is only about 10% of its original size (Frank et al, 2011).

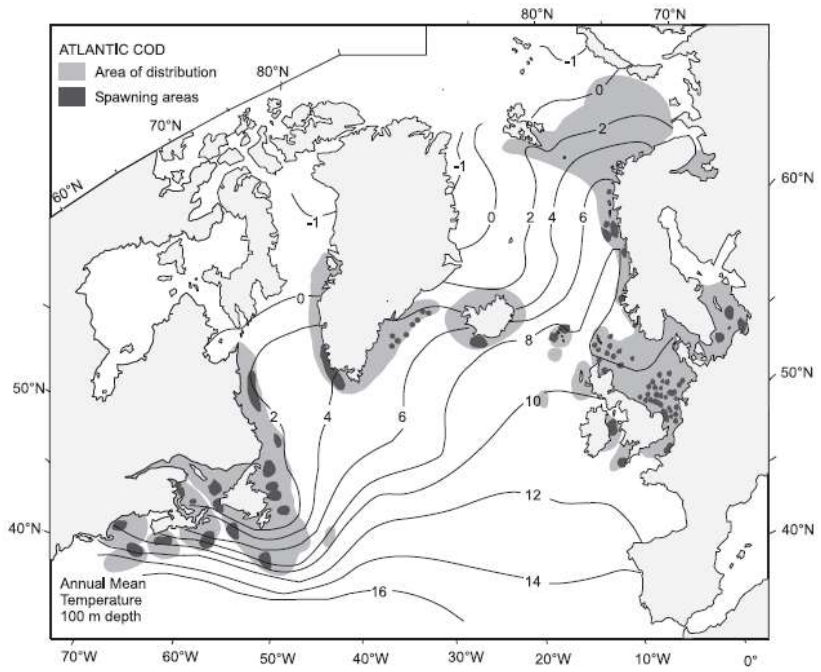

Figure 1.1 Cods distribution. *Distribution and spawning areas of cod in the North Atlantic, and the temperature of the waters (Sundby, 2000).*

1.3.2 Faroese cod populations

Atlantic cod is split into several populations, two of which are in the Faroese waters, one on the Faroe Shelf and the other on Faroe Bank. Studies have shown that cod from Faroe Bank grows faster than cod in any other known population (Figure 1.2). Three year old cod from Faroe Bank is on average 73 cm in length and weighs 4.9 kg, where as cod from the Faroe Shelf is just 55 cm in length and weighs 1.7 kg (Magnussen, 2007).

The fast growth of cod from Faroe Bank, makes it interesting for aquacultures since there the cod can be produced in a controlled manner. Cod hatcheries would want their eggs to come from a population of cod that grows faster than other cod. This begs the question**,** why does the cod on from Faroe Bank grow faster than cod in other waters. Two possible reasons are either they have "superior" growth genes, or it could be environmental factors, as food,

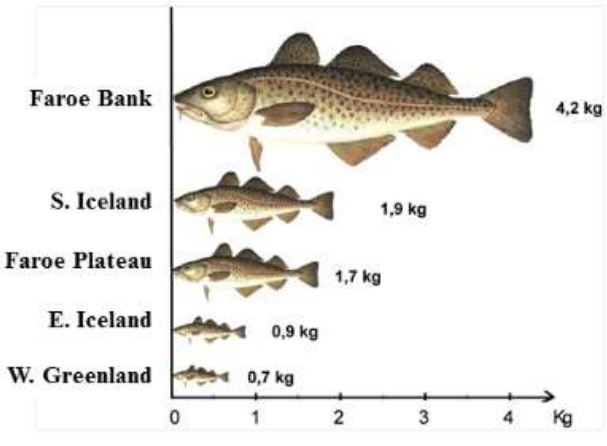

Figure 1.2 Weight of three year old cod. *Weight of three year old cod from various locations in the North Atlantic (Magnussen, 2009).*

temperature or other environmental factors. It could also be a combination of the genes and the environment. Selling cod from Faroe Bank as a genetic template is interesting if its genetic code is the factor that makes it grow faster than other cod. If environmental factors are the reason, then it would be interesting to know what these factors are, and if these environmental factors could be replicated for aquacultures.

The environmental questions are questions for another project though, as this project deals with the genetics of haemoglobin.

1.4 Haemoglobin

1.4.1 Globin proteins and haemoglobin

Globins are expressed in various tissues and cell types, where the most commonly known ones are myoglobin in muscles, cytoglobin in brain tissue and haemoglobin in red blood cells (Hoffman et al, 2012).

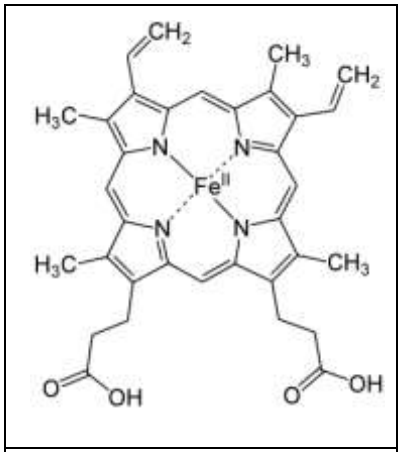

Figure 1.3 Haeme structure. *Structure of Haeme group in Haemoglobin and how the haeme has iron tied in a heterocyclic ring.*

Globin proteins can be single proteins, like myoglobin which facilitates oxygen diffusion and can also function as an oxygen reservoir for *Cetaceans* and *Pinnipeds* (Antonini, 1971). Globins can also be multimeric complexes, like the tetrameric haemoglobin (Hoffman et al, 2012). In the gnathostomes, the haemoglobin consists of four subunits proteins, with two identical α-subunits and two identical β-subunits. The haeme groups are noncovalently bound to the haemoglobin molecule. The haeme group has iron tied in a heterocyclic ring (Figure 1.3). Each globin peptide chain binds one haeme group, which means that each haemoglobin can carry four molecules of oxygen (Alberts et al., 2009).

The four subunits are chained together around the haeme parts (Figure 1.4). The oxygen binding sites in all the haemoglobin interact which allows an allosteric change in the molecule, such that it binds and releases four oxygen molecules at a time. This feat becomes very important in larger animals, because with increasing size, it becomes increasingly important to be able to transport oxygen throughout the body, as direct diffusion of oxygen from the air to all cells becomes less and less of an option with increasing size (Alberts et al., 2009).

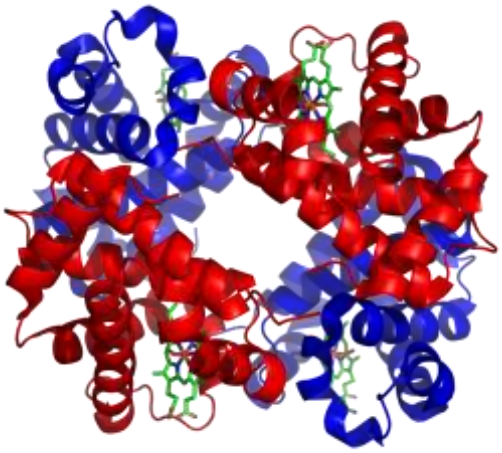

Figure 1.4 Human haemoglobin. *The red parts of the protein are the two α subunits, and the blue parts are the two β subunits. The four green groups are the iron with the haeme group.*

Not all haemes have the biological function of transporting oxygen. Their function is to transport diatomic gasses, chemical catalysis, diatomic gas detection and electron transfer. It is speculated that the original evolutionary function of haeme was electron transfer in primitive sulphur based photosynthesis pathways in ancestral cyanobacteria before the appearance of molecular oxygen (Hardison, 1999).

1.4.2 Globin gene evolution

Globin genes have some similarities, like the highly conserved globin fold, which is a three dimensional folding of six to eight α-helices, that enclose the haeme group in a hydrophobic pocket (Hoffmann et al., 2012). The similarities between globins in higher vertebrates, marine worms and insects show that they all must derive from a common ancestral gene. This original monomeric globin protein transported oxygen molecules throughout the animal's body (Alberts et al., 2009). The α and β globins probably arose from a duplication followed by mutation, of the original globin gene about 500-570 million years ago (Figure 1.5) (Alberts et al., 2009). As the fishes branched off from other vertebrates after α and β globins had evolved, they have both α and β haemoglobins. Both α and β haemoglobin have since then had several more

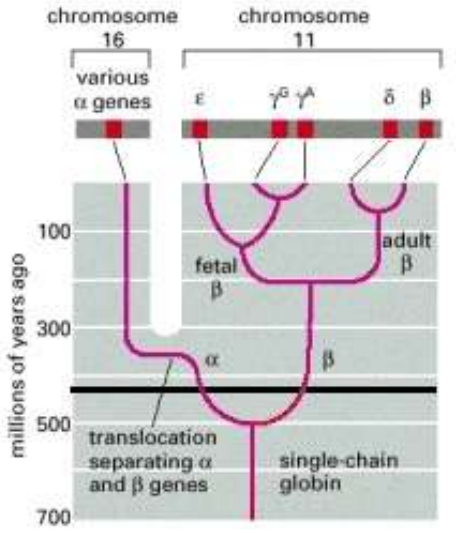

Figure 1.5 Evolution of human haemoglobin. *Repeated rounds of duplications and mutations generated the globin gene family (Alberts et al., 2009). Actinopterygii, where Teleostei evolved from, branched off from other vertebrates about 420 million years ago, indicated by the black line (Märss, 2001).*

duplications followed by mutations which has given rise to the nine different haemoglobin genes in cod studied in this project.

1.4.3 Globin genes: Family and gene structure

As for most eukaryotic genes, also the globin genes are split into exons and introns. Exons are the coding sequences of DNA. Intron are non-coding DNA sequences that are transcribed, but then cut in the transcription of mRNA (Hartl and Jones, 2006).

Nine different haemoglobin genes have been discovered in cod (Borza et al., 2009). Most of them have several alleles (Table 1.1). Some of the genes are more similar to each other than others (Figure 1.6) (Table 1.2) for example β3 and β4 only differ from each other on seven nucleotides, β3 has a chain of four additional nucleotides in an intron, resulting in only one amino acid difference between β3 and β4. Table 1.2 clearly shows how some of the genes have more common characteristics than others and most clearly β3 and β4, and it supports the data from Figure 1.6.

Table 1.1 Haemoglobin alleles. *Number of alleles of cod haemoglobin (Genbank)*

Gene	Number known of alleles according to Borza et al. (2009)
α-1	2
α-2	1
α-3	3
α-4	2
β-1	8
β-2	2
β-3	3
β-4	4
β-5	2

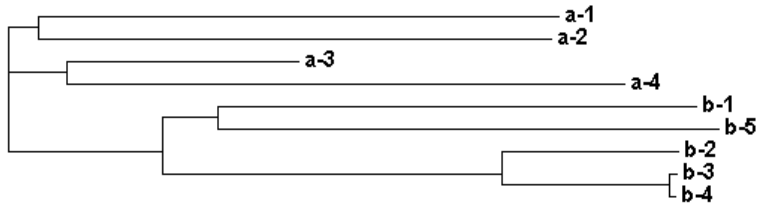

Figure 1.6 Phylogram tree of cod haemoglobin genes. *Phylogram tree of α1- α4 and β1- β5 DNA sequences compared against each other. The full-lenght gene sequences were paired against each other with ClustalW.*

Table 1.2 All genes compared to each other.
One allele of all nine genes aligned with clustalw. These numbers show percentage identity.

	a1-1	a2-1	a3-1	a4-1	b1-1	b2-1	b3-1	b4-1	b5-1
a1-1	100	65.36	73.07	60.72	59.95	59.4	56.19	56.92	56.28
a2-1	65.36	100	70.53	61.55	62.24	60.13	59.81	59.16	55.86
a3-1	73.07	70.53	100	73.44	60.58	65.75	74.27	73.85	70.07
a4-1	60.72	61.55	73.44	100	59.75	58.37	52.99	52.45	53.53
b1-1	59.95	62.24	60.28	59.75	100	59.1	66.62	67.27	67.17
b2-1	59.4	60.13	65.75	58.37	59.1	100	88.32	88.25	69.1
b3-1	56.19	59.81	74.27	52.99	66.62	88.32	100	99.58	62.17
b4-1	56.92	59.16	73.85	52.45	67.27	88.25	99.58	100	62.8
b5-1	56.28	55.86	70.07	53.53	67.17	69.1	62.17	62.8	100

1.4.4 Haemoglobin as a genetic marker in cod

Table 1.3 Weight and isoeletric point of genes and alleles. *Haemoglobin sequences were put into MS-digest tool at prospector.ucsf.edu and the tool calculated the molecular weight and isoelectric point of the protein*

Gene/alleles	MW (unmodified)	pI
a1-1,2,e	15918	9.4
a2	15797	8.9
a3-1,2,3	15967	9.2
a3-e	15983	9.2
a4-1,2	16041	7.8
b1-1,2,5 (β1A)	16373	7.7
b1-6 (β1A)	16391	7.7
b1-3,8 (β1B)	16302	7.0
b1-4,7,e (β1B)	16284	7.0
b2-1,2	16797	5.9
b3-1,2,e	16635	6.2
b3-3	16662	6.2
b4-1,2,3	16622	6.2
b4-4	16636	6.2
b5-1,2	16738	5.6

Haemoglobin can be used as a population marker, and was among the early population genetic markers used for cod. Sick found in 1965 that there are three phenotypic types of haemoglobin Hb-I (1/1), Hb-I (2/2) and Hb-I (1/2) as resolved by agarose gel electrophoresis. These three phenotypes are found in different frequencies in various environments (Table 1.4). Hb-I (1/1) is most commonly found in warmer regions, Hb-I (2/2) is most commonly found in colder regions and Hb-I (1/2) is found in between (Sick, 1965).

Agarose gel electrophoresis was among the first molecular methods, but is insensitive compared to other methods, and is but it hardly used today. Isoelectric focusing allows additional separation compared to gel electrophoresis (Borza et al. 2009), and can be used as an optional step in electrophoresis or independently of electrophoresis.

Table 1.4 Allele frequencies. *Haemoglobin 1 allele frequencies for cod in the North Atlantic ocean (Magnussen, 1993)*

Location		Allele frequency
Norwegian cod	Barents Sea	0.03 – 0.21
	Northern coastal cod	0.10 – 0.41
	Southern coastal cod	0.36 – 0.52
Eastern Baltic Sea		0.01 – 0.05
Danish sounds		0.57 – 0.64
North Sea		0.55 – 0.72
Faroe Islands		0.04 – 0.10
Greenland and Iceland		0.01 – 0.02
North America	Newfoundland	0.04
	Gulf of Maine	0.07

All of the nine different haemoglobin genes are claimed to be expressed simultaneously in adult cod, but the level of the expression varies (Borza et al, 2009). β1 gene appears to be the gene that sets the phenotype, individuals with phenotype Hb-I (1/1) only had the β1A variation, and individuals with Hb-I (2/2) only had β1B variation. Individuals with Hb-I (1/2) had both variations of β1. The difference between β1A and β1B is two non-silent mutations that change the amino sequence, β1A has Met55 and Lys62 where β1B has Val55 and Ala62 (Table 1.3) (Borza et al., 2009).

1.4.5 Haemoglobin as a potential development and environmental marker or sensor

Cod will adapt to the water temperature it occupies, which will favour the expression of more haemoglobin components. This will also affect the oxygen affinity of haemoglobin (Brix et al., 2004).

It has been shown that the haemoglobin phenotype changes as the juveniles grow up. Juveniles with Hb-I (1/1) preferred an ambient temperature of 15.4^{o}C where as juveniles with Hb-I (2/2) preferred an ambient temperature of 8.2^{o}C (Petersen and Steffensen, 2003). For fish, the preferred temperature is closely associated with the optimal growth temperature (Fry and Hochachka, 1970).

Similarly to other fish species such as rainbow trout, carp and goldfish, cod could possibly be able to respond to environmental challenges, like chronic hypoxia or long term changes in temperature, by altering the expression level of each of the different genes, thereby adapting itself to a new environment (Borza et al., 2009).

1.5 Aims of present project

The aims of the present project is studying the haemoglobin genes and alleles, which will give us more detailed data about the Faroese cod population and its subpopulations, and can give us an overview of haemoglobin alleles in the Faroese waters.

Because this institution does not have any history in molecular DNA methods, this project aims to establish the methods leading up to large-scale parallel sequencing of cod. These methods include primer design and the testing of primers and optimizing conditions for PCR.

2. Materials and methods

2.1 Sampling

Blood samples were taken from cod at the Faroese Aquaculture Research Station where they study cod in various projects, with individuals from Faroe Bank and the Faroe shelf. 64 samples were taken from Faroe Bank cod and 28 samples from Faroe Shelf cod. The sampling procedure was: About 100 µL of blood was collected by a heparinized syringe and centrifuged with 500 µL Tris buffer with 25 units of heparin/ml. The supernatant fluids were discarded, and additional 500 µL Tris was added and centrifuged again. After the supernatant fluids had been discarded again 250 µL RNAlater was added and the tube was gently vortexed to mix RNAlater well. Because the samples needed to be stored over longer periods, they were first put into the refrigerator for a couple of days, to let RNAlater penetrate into all the cells, after a few days samples were moved to the freezer.

2.2 DNA purification

For purification of DNA from the cod blood, "DNeasy® Blood & Tissue Kit" by Qiagen was used according to the supplier's procedure. The optional RNAse treatment was not used.
The concentration of purified DNA was estimated by NanoPhotometer Pearl from Implen.

2.3 Primers

The primers (Table 2.1) were manually designed so that they could cover as much of the gene as possible. Primer pairs were located to give PCR products that were of similar lengths, for cDNA about 350 and for gDNA 550-700 base pairs (Figure 2.1). Some of the Hb genes could be covered by one primer pair, where other Hb genes needed two or even three primer pairs to cover the entire gene. The primers were all designed to work optimally at similar temperatures, so that it would be possible to run many combinations under identical conditions. Accession numbers for all the haemoglobin genes and their alleles are found in Table 2.2.
All standard PCR primers were synthesized by DNA Technology (Aarhus), and the 454 sequencing PCR primers were synthesized by either DNA Technology or TAG Copenhagen.

Table 2.1 Primers. *All the primers used in this project.*
Some primers accounted for degenerated sequences.' W = A and T,''M = A and C

Primer name	Sequence
a1F14	CAAAGGACAAGGCGACCGTCA
a2F-27	CTAGAAAGCAACTATCTGAACGTCAAC
a2F312	CTCTCACTGGAAGGACGCGAG
a2F352	AGGAAGCATGGCATCACCATCATG
a3F59	TGGCTGACCGCATAGGAGCAG
a3F851	ACCTCGCTCCTCTTCAAACCTAC
a3F1163	GGTTAACTAGGACCATAGGTGTAC
a4F63	AGCAGTGGAGATCGGACACCAG
a4F249	CCAAACCAAAGTCTACTTCTCCCA
b1F-26	CATTGAACCCTTAAGACWACGCCAC'
b1F-44	GAGATTCAGCCTAAGCTACATTGAAC
b1F44	TCTGGAGCAAGATCGACATTGATG
b1F589	CACTCCGAGAAACTGCACGTCGA
b1F-35	CCTAAGCTACATTGAACCCTTAAGAC
b234F-13	CACAWCAGCAACCATGGTTGAG'
b234F255	TCTGATCGTGTACCCCTGGACC
b234F50	CAMCTTGGACTACGAAGAGATCGG''
b34F-5	CACCATGGTTGAGTGGACAGATARTG
b5F1	ATGGTAGAGTGGACAGAATTTGAG
b5F27	CGATACAATCAAGGACATCTTCTC
b5F239	ACTCGGAGGTACTTTGGAAACTTC
a1R400	AAGCACACAGGAACTTGTCGAC
a1R414	AGCTCGCTCAGAGTGAGAAGAC
a2R346	GGAAGATCATGGACATGCACACCA
a3R394	ACGCAGAGAGATACTTGTCCATG
a3R1304	CCTATACCTATCTGTCTAGTACAC
a4R614	CTGAGGAGTGAAATCATCTGGGAAC
b1R252	GGACTTGATTTCGTCCATGTGGTC
b1R463b	TAGCGTCAGCACACATCTTCTAGTG
b1R463	TAGCGTCAGCACACATCTTCTAG
b1R611	CACTCCGAGAAACTGCACGTCGA
b1R261	GACGAAATCAAGTCCACCTACGCT
b234R417	GACGACGACAGACAGGAACTTCTG
b234R447	CTTGTCAGAGTGGAGCAGAGAC
b2R525	AGATCACCCACTCTGCGGCTTG
b5R449	CGCTACTAGTGGTACTGGCTTC
b5R709	ACGGTAGTCCTACTGTCCATGCA

Table 2.2 Hb genes on Genbank.
Accession numbers for the haemoglobin genes in Genbank.

α1	FJ666964 - FJ666965
α2	FJ666966
α3	FJ666967 - FJ666969
α4	FJ666970 - FJ666971
β1	FJ666972 - FJ666979
β2	FJ666980 - FJ666981
β3	FJ666982 - FJ666984
β4	FJ666985 - FJ666988
β5	FJ666989 - FJ666990

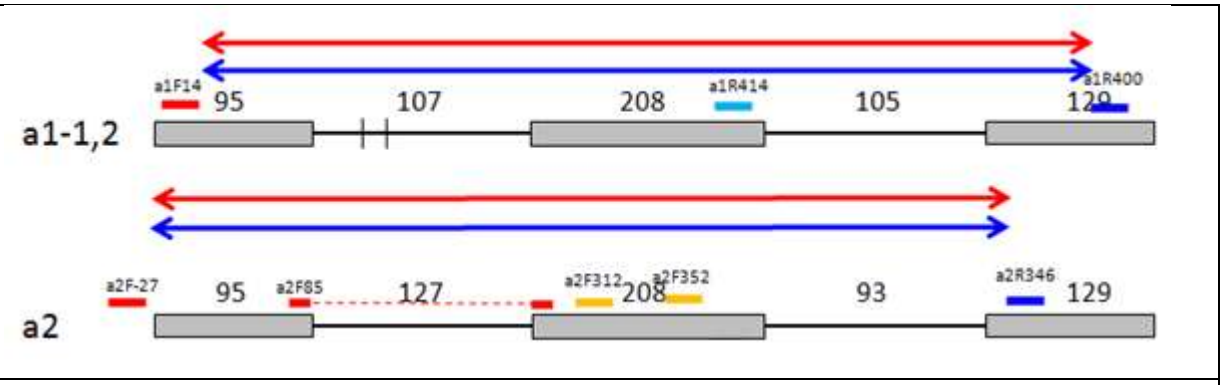

Figure 2.1 Primer locations for α1 and α2. *Illustration showing location for primers on the haemoglobin gen. The exons are the bulging boxes and the introns are the lines between the boxes. The sizes of the exons and introns are indicated by the numbers above them. The horizontal red arrow indicates the primer is for cDNA and the blue arrow is for gDNA. The red boxes indicate a forward primer and the blue boxes indicate reverse. The yellow and light blue are newer primers. The vertical lines show positions of SNPs generating different alleles.*

2.4 PCR

Two thermostable DNA polymerases were used during this work, a Platinum Taq DNA Polymerase (Invitrogen) and HotStar HiFidelity DNA Polymerase kit (Qiagen). Taq was used for the main part of this work.

PCR with Taq followed the standard protocol procedure from Invitrogen with the final primer concentration 0.2 µM and DNA about 150 ng.

The PCR machine was setup according to the procedure for Taq. Denaturation was for 15 sec at 94 ^{o}C, annealing for 1 min at 60 ^{o}C and elongation for 1 min at 72 ^{o}C. The PCR ran for 34-40 cycles, depending on how much of the PCR product we wanted. After all the regular cycles, an extra step with 10 min elongation is performed.

For next generation sequencing it is recommended to use a highfidelity DNA polymerase. For this HotStar HiFidelity DNA Polymerase (Qiagen) was used. This kit needed some optimizations of conditions before any results were acquired. This optimization is shortly described in Results.

2.5 Electrophoresis

2.5.1 Polyacrylamide gel electrophoresis

A 5% polyacrylamide gel was used. The liquid acrylamide was 30% acrylamide / Bis solution 37.5:1 (2.6% C) from Bio-Rad. TAE was used as the gel buffer as well as electrophoresis buffer. 1x solution of TAE contains 40 mM Tris, 20 mM acetic acid, 1 mM EDTA and has pH 8.3.

The gels stood overnight before use.

2.5.2 Staining of polyacrylamide gels

Silver staining was used to stain the DNA in the PAGE gels.

The steps were

1. Put the gel on a rocking table with 2 µg/mL sodium thiosulfate for 25 min. For the box containing my gel about 50 mL was needed.
2. Rinse with distilled water.
3. Put on a rocking table with 1 mg/mL $AgNO_3$ for 20min.
4. Two short rinses with distilled water.
5. Development in 3% sodium carbonate (Na_2CO_3) + 1.5 µL formalin (37%) / mL
6. Stop development with 10% acetic acid.

2.5.3 Agarose gel

The molding form for the gel was 10 cm x 6.2 cm. Gel thickness was 0.5 cm. For a 1.5% gel 0.4838 g agarose powder was mixed with 31 mL TAE.

2.5.4 Fluorescent staining of agarose gels

The UV dye (Biotium GelRed nucleic acid stain 10.000x in water) was mixed 1:10.000 with the agarose. The gels were illuminated at 254 nm to visualize the dye complexed with DNA, and images were captured with a handheld digital camera.

2.6 Purification of DNA from agarose gels

JETquick Spin Columns were used, and the supplier's procedure was followed exactly. For the elution 50 µL of sterile water was used.

2.7 Dideoxy sequencing

2.7.1 Sequencing reaction

QiaQuick PCR purification kit (Qiagen) was used for purification of the PCR products before the sequencing reaction. The sequencing reaction was performed with Big Dye 1.1 (Applied Biosystems), and the supplier's procedure was used. The sequence reaction was then run in the PCR machine with these cycle settings: 96^{o}C for 1 min, then 30 cycles of: 96 ^{o}C for 10 sec. 50 ^{o}C for 5 sec and 60 ^{o}C for 4 min.

2.7.2 Purification of sequencing products

The procedure was obtained from Debes H. Christiansen of FVA (Food and Veterinary Agency, Tórshavn), and was used with some minor modifications:

1: 1 µL 3M NaAcetate (pH 4.6) and 25 µL 96% EtOH were added to the PCR tube with 10 µL of sequencing products. This was mixed, and can then stand in room temperature for 15 min - 24 hours, we used 30 min.

2: The PCR tube was centrifuged at 20.000g for 15 min.

3: All the liquid in the PCR tube was discarded, and then 80 µL of 70% EtOH was added.

4: All liquid was discarded and the PCR tubes dried in the vacuum drier at room temperature for 30 min.

After these steps the samples were ready to be sent to FVA to be sequenced.

2.7.3 Sequencing

The products of the sequencing reaction were analyzed on an ABI 3100 by Debes H. Christiansen from FVA.

3 Results

3.1 Steps towards parallel sequencing of targeted genes

The overall aim in the project, was to establish usable primers and parts of the methods needed to perform large-scale sequencing of targeted genes by parallel sequencing, also called “next generation” sequencing. The targeting itself is performed by PCR. Thus, the steps that need to be established are: Suitable design of primers and their locations, testing of primers and optimizing PCR conditions, optimizing the purifying procedure, mixing of PCR products to make suitable substrate for the large-scale parallel sequencing (in this case 454 sequencing), the bioinformatics analyses of the obtained sequences and finally the interpretation of the sequencing results.

The focus in the present part of the project, however, was the testing of primers, and this will be described in some detail. Our institution does not have any history of working of molecular DNA methods. Thus, parts of the methods described here were established in our laboratory during this project.

3.2 Polyacrylamide and agarose gels and their staining:

I have switched between using Polyacrylamid gel electrophoresis (PAGE) and agarose electrophoresis, this because of the different properties of the two methods. PAGE has advantages in that can give better separation between bands that are close and the bands are usually sharper than in agarose. Silver staining of PAGE gels is sensitive and easy to perform, but it is time consuming. In agarose gels, you can easily cut out the individual bands and purify the PCR product for subsequent procedures, for example, a second round of PCR. This is ideal if the primers are not fully specific for one particular gene, and you therefore get several bands in the gel. Agarose is safer to handle than polyacrylamide in powder form. If fluorescent staining is used, the staining is a lot easier and faster with agarose gels.

In this project I have used two different methods of staining the gels. For PAGE I have used silver staining. This is a pretty simple method, it is time consuming and you run the risk of damaging the gel, because they are so thin and easy to rip.

Staining of the agarose gels was done with a fluorescent dye, GelRed, this method is incredibly easy and fast (Figure 3.1). This is because the fluorescence dye can be added during the agarose gel molding, just after it has been heated. I have found this staining method to also provide the agarose gels with very good separation between the bands. It is harder to take a good picture of the florescent stained gel, at present we must use a handheld camera with sufficient light sensitivity that can be manually focused. For the PAGE gels I have found the camera on my cell phone to work very well. It is worth mentioning that the separation between bands was clearer when seen in the lab, where you could easily see all the bands on the size ruler.

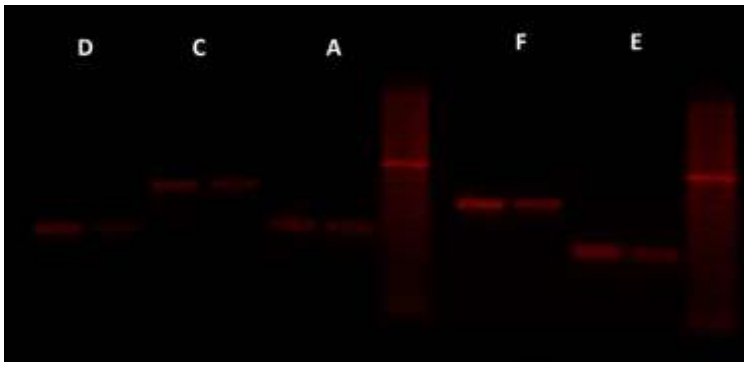

Figure 3.1 *Testing of the UV dye. 5x more PCR product was loaded in the samples on the left in pairs D, C, etc.*

3.3 Taq vs. HotStar:

In the initial experiments with HotStar Hifidelity DNA Polymerase from Qiagen, the experiments were performed according to the supplier's recommendations, but very high background was found (Figure 3.2). A few experiments were done to resolve the reasons for the high background, but with limited success. Due to the little time available, we have mainly used Taq in the experiments described here. Subsequent experiments done by other collaborators have indicated that less enzyme, less primer and preferably low amounts of DNA should be used with HotStar.

Figure 3.2 *Non optimized tests with HotStar, R is the molecular ruler.*

3.4 Testing of standard primers:

3.4.1 Primers for haemoglobin α genes:

In this project we started out with testing of haemoglobin primers. These had all been designed from the known haemoglobin genes found in Genbank. The primers for the alpha genes proved to generally work great, and provided us with very good and clean bands, that all were in the expected positions. Some of the primers lacked specificity, as described below for the different Hb genes. This was indicated by several bands. Some of the bands were not of the lengths corresponding to any known allele that the primer pair was expected to recognize. Other primer combinations just did not provide any result at all. To try to fix both these problems, we tried to setup a temperature gradient, but this showed to diminish the wanted PCR products at close to the same rate as the unwanted ones (Figure 3.3).

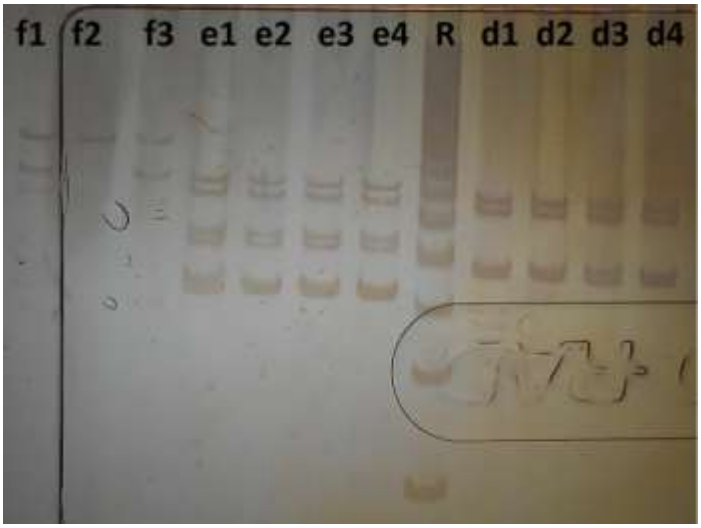

Figure 3.3. *Temperature gradient of some primer combinations*. f: b34F-5 and b234R417, e: b234F50 and b234R447, d: b34F-5 and b234R447. The gradient is 1:59, 2: 59.8, 3: 61.4 and 4: 63. R: molecular ruler. Expected products were f: 646 and 652; e: 346, 353 and 349; d: 408 and 414.

For the overview of the results of the primer combinations, see appendix A.

3.4.1.1 PCR targeting and dideoxy sequencing of Hba1 PCR products

The primers for Hba1 functioned very well, and did not pose any difficulties, and gave clean bands at the expected location in electrophoresis (Figure 3.4) (Table 3.1)

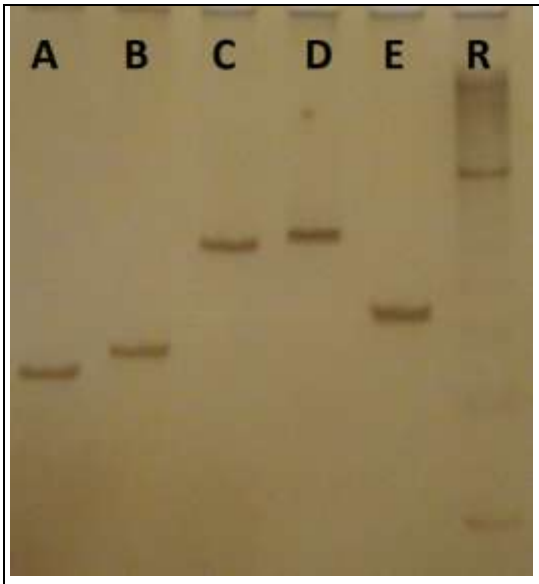

Figure 3.4 α1 and α2 primers. *A = a2F352 and a2R346, B= a2F312 and a2R346, C = a2F-27 and a2R346, D = a1F14 and a1R400, E = a1F14 and a1R414, R = molecular ruler.*

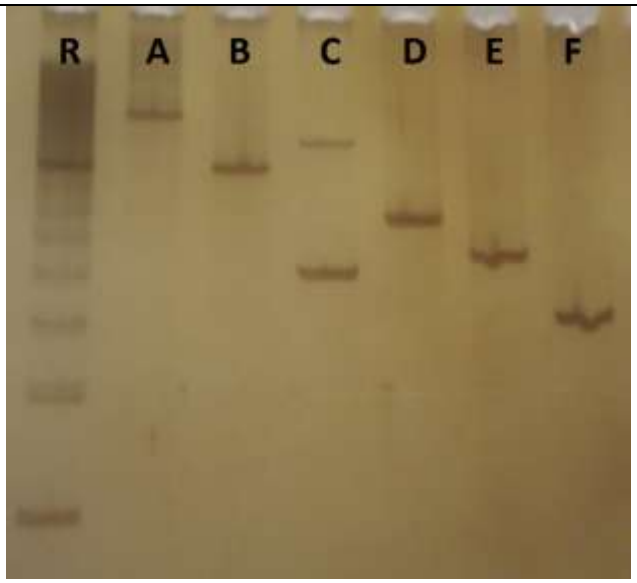

Figure 3.5 α3 and α4 primers. *A = a3F59 and a3R394, B= a3F851 and a3R394, C = a3F851 and a3R1304, D = a3F1163 and a3R394, E = a4F63 and a4R616, F = a4F249 and a4R616.*

Table 3.1 *Primer combinations for Hba1 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
a1F14	a1R400	599	570
a1F14	a1R414	356	330

The sequencing of α gave very clean sequence that matches well with the expected results (Figure 3.6). However the three positions in Figure 3.6 that our sequence differs from the published ones hint at different conclusions, Position 113 differs from both published alleles, Pos. 116 is the same as Gm-a1-2 and Pos. 122 is the same as Gm-a1-1. This could indicate that our sequence is a new allele, but more detailed investigations would be needed to confirm this suggestion.

```
Gm-a1-1          AAGGCCGAGCTTATCGGCGCCGATGCTCTTTCAAGGTATGCTTACTCCAGCCACACAATT 120
Gm-a1-2          AAGGCCGAGCTTATCGGCGCCGATGCTCTTTCAAGGTATGCTTACTCCAGCCACAAAATT 120
2                AAGGCCGAGCTTATCGGCGCCGATGCTCTTTCAAGGTATGCTTACTCCAGCC-CAAAATT 85
                 ************************************************** ** ****

Gm-a1-1          ACAGTTAGGGGGAATAACTTGAACAATTTGCTGTGTTTTTTTTATTTATTAATGTATTTA 180
Gm-a1-2          AGAGTTAGGGGGAATAACTTGAACAATTTGCTGTGTTTTTTTTATTTATTAATGTATTTA 180
2                ACAGTTAGGGGGAATAACTTGAACAATTTGCTGTGTTTTTTTTATTTATTAATGTATTTA 145
                 * ************************************************************
```

Figure 3.6 Alignment of the sequence from the PCR product of the primer pair a1F14 + a1R400. *Gm-a1-1 and Gm-a1-2 are the two published alleles of Hba1, the sequence labelelled 2 is from our PCR product. a1F14 used as sequencing primer. The area marked in purple is the 3'-end of exon 1, followed by intron 1.*

3.4.1.2 PCR targeting and dideoxy sequencing of Hba2 PCR products

The primers for Hba2 functioned very well, and gave clean bands at the expected location in electrophoresis (Figure 3.4) (Table 3.2)

Table 3.2 *Primer combinations for Hba2 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
a2F-27	a2R346	592	550
a2F85	a2R346	482	470

```
Gm-a2           TCAGCCTTAGCGAGCTGCACGCGTTCATGCTGAGAGTTGACCCCGTCAACTTCAAGCTCC 360
3               TCAGCCTTAGCGAGCTGCACGCGTTCATGCTGAGAGTTGACCCCGTCAACTTCAAGCTCC 307
5               TCAGCCTTAGCGAGCYGCACGCGTTCATGCTGAGAGTTGACCCCGTCAACTTCAAGCTCC 179
                *************** ****************************************
```

Figure 3.7 Alignment of the sequences from PCR products from the primer pairs for α2. *Gm-a2 is the published sequence of Hba2. Sequence 3 is a2F-27 + a2R346. Sequence 5 is a2F85 + a2R346. Both 3 and 5 use the forward primer as sequencing primer. Area marked in purple is the 3'-end of exon 3. Nucleotide Y is both C and T.*

The sequencing most likely shows us the only known allele of α2 (Figure 3.7), the Y (C and T) base marked in teal is the only difference in the sequence, and it may be a sequencing error, since we only see it on sample 5 and not 3. Further investigations are needed to say anything with certainty.

3.4.1.3 PCR targeting and dideoxy sequencing of Hba3 PCR products

Most primer combinations functioned well, and gave the expected lenghts of the products. In sequencing, the submitted product from the primer pair of a3F59 and a3R394 only resulted in a 340 nucleotides long readable part of the sequence.

Table 3.3 *Primer combinations for Hba3 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
a3F59	a3R394	1701	1600
a3F851	a3R394	953	1000
a3F851	a3R1304	453	430 + 1200
a3F1163	a3R394	640	600

```
Gm-a3-1          GGAAAA-ATCGAAGGTAAATTAACCGAAGGCACACATAAATTGAATCTGGTAACAATGTT 1679
Gm-a3-2          GGAAAA-ATCGAAGGTAAATTAACCGAAGGCACACATAAATTGAATCTGGTAACAATGTT 1679
Gm-a3-3          GGAAAA-ATCGAAGGTAAATTAACCGAAGGCACACATAAATTGAATCTGGTAACAATGTT 1679
10invkompl       GGAAAACKTCGAAGGTAAATTAACCGAAGGCACACATAAATTGAATCTGGTAACAATGTT 308
                 ******  **************************************************
```

Figure 3.8 Alignment of the sequences from PCR product of the primer pair a3F59 + a3R394. *Gm-a3-1, Gm-a3-2 and Gm-a3-3 are the three published sequences for Hba3. a3R394 used as sequencing primer. K which is marked in teal is G and T.*

Most of the α3 sequence was unreadable, but the 340 nucleotides that were readable had high quality readout. Our obtained sequence was located outside the area where the published alleles differ (Figure 3.8). The only disparity, marked in teal, from the three known alleles is most likely a sequencing error or possibly a new allele. New experiments are needed to determine which alternative is the correct one.

3.4.1.4 PCR targeting Hba4

Both primer combinations for α4 worked fine in electrophoresis (Table 3.4) (Figure 3.5), but both failed in sequencing, and I have no sequencing results to show for α4.

Table 3.4 *Primer combinations for Hba4 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
a4F63	a4R614	543	460
a4F249	a4R614	365	320

3.4.2 Primer for haemoglobin β genes

3.4.2.1 PCR targeting and dideoxy sequencing of Hbb1 PCR products

There were many primer combinations tested for β1, where some worked good, and others did not work at all (Table 3.5).

Table 3.5 *Primer combinations for Hbb1 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
b1F-26	b1R252	544	500 smear 530 clear
b1F-26	b1R463b	1182	blank
b1F-26	b1R463	1182	blank
b1F-26	b1R611	584	580 600 630
b1F-26	b1R261	530	380 400

			420
b1F-44	b1R252	562	540
b1F-44	b1R463b	1200	Blank
b1F-44	b1R463	1200	Blank
b1F-44	b1R611	609	220 580 600 630
b1F-44	b1R261	571	220 380 400 420
b1F44	b1R252	475	Blank
b1F44	b1R463b	1113	Blank
b1F44	b1R463	1113	Blank
b1F44	b1R611	522	Blank
b1F44	b1R261	485	Blank
b1F589	b1R463b	614	600
b1F589	b1R463	614	650
b1F-35	b1R463b	1191	Blank
b1F-35	b1R463	1191	Blank
b1F-35	b1R611	600	580 600 630
b1F-35	b1R252	552	530
b1F-35	b1R261	563	380 400 420

Some of the primer combinations resulted in several bands. All bands from the primer pairs b1F-44 + b1R611 and b1F-44 + b1R261 were purified and subjected to a new round of PCR with the same primers (Table 3.6).

Table 3.6 Second PCR on select samples. *Sample K is the primer pair b1F-44 + b1R611 and sample U is the primer pair b1F-44 + b1R261. They needed to be renamed for the second PCR, because there were four new PCR products from each of the two original. Where the original naming was alphabetical, the new letter added ment: x = extra small, s = small, m = medium, l = large.*

	K	U	New name:
Observed results:	220 580 600 630	220 380 400 420	Kx – Ux Ks – Us Km – Um Kl – Ul
Expected results:	609	571	

The second PCR did not successfully isolate the bands.

The obtained sequences from Kx and Ux pointed out the same area of Hbb1. Kx stops in mid exon 2 (Figure 3.9) and Ux aswell stops in the exact same position, the area that both Kx and Ux skip does have an area where there are alot of A and T. Kx and Ux are may be the products of a pseudo-gene, but it is more likely that there is some kind of secondary structure (in spite of the high temperature) has caused the polymerase to skip the intermediate area in Hbb1. It is interesting to note that for both K and U there is a T (Figure 3.8 marked in teal) where a C was expected, and this suggests that there could be a new allele.

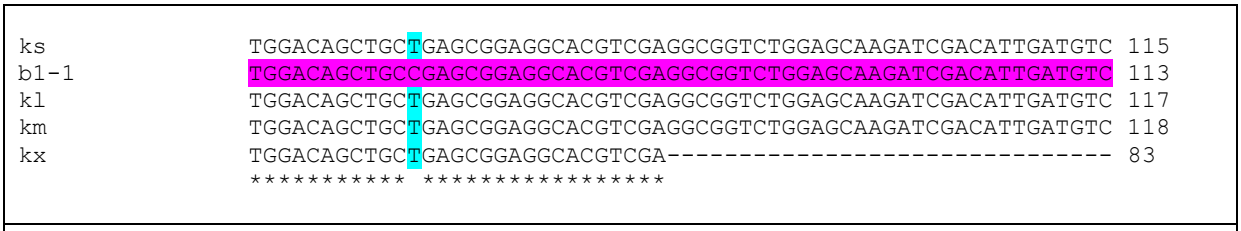

Figure 3.9 Alignment of the sequences for PCR products from the primer pair b1F-44 + b1R611. *b1R611 used as sequencing primer. Sequencing of several bands from the same primer combination sequencing primer. Exon is marked in purple, and teal is where the sequence differs from Genbank sequence.*

To optimize alignment results for the sequences from b1F-44 and b1R611, Kx was removed (Figure 3.10).

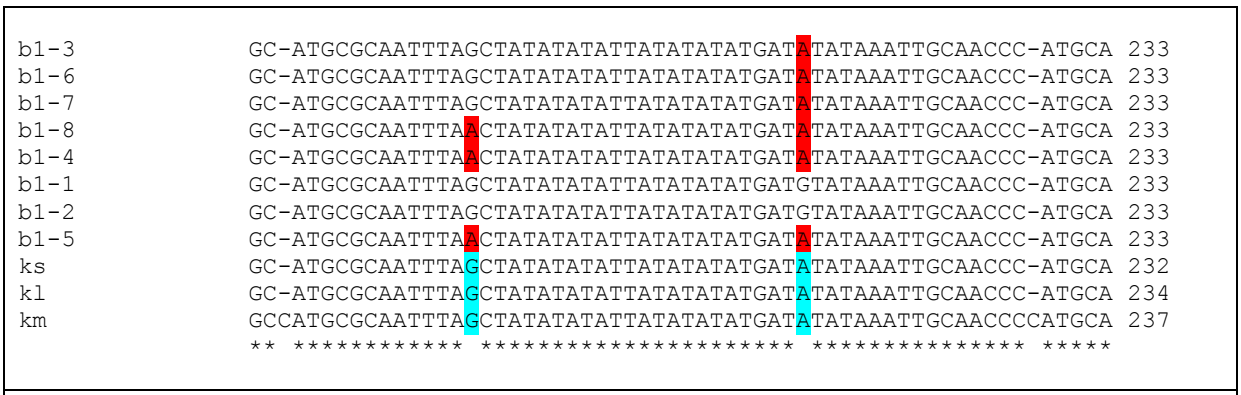

Figure 3.10 Kl, Km and Ks paired with all eight β1 alleles. *Teal marks the areas where our sequence products differs from some of the known alleles, and red marks where the known alleles differ from each other. b1-3, b1-6 and b1-7 match perfectly except from the extra C's in Km. The sequence is part of intron 1.*

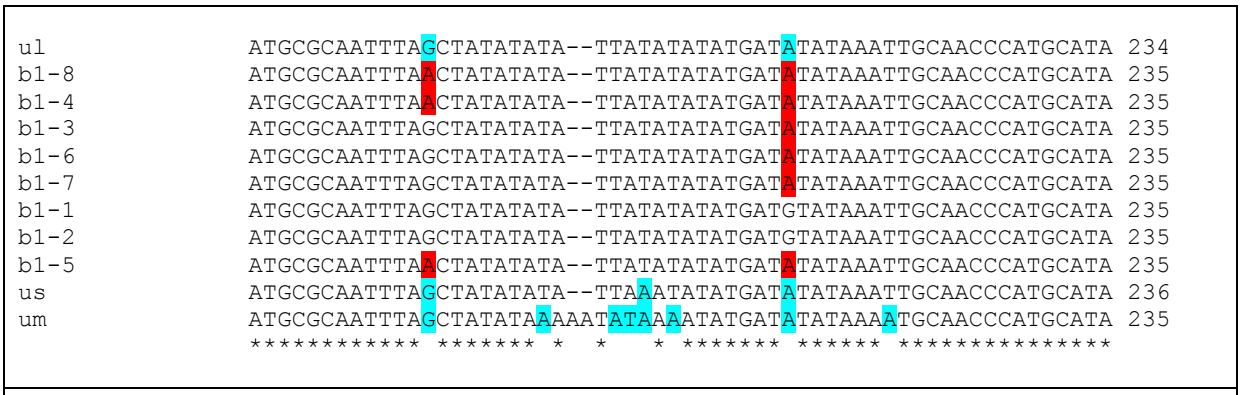

Figure 3.11 Alignment of the sequences for PCR products from the primer pair b1F-44 + b1R261. Ul,Um and Us paired with all eight β1 alleles. *Teal marks the areas where the sequence differs from the known alleles, and red marks where the known alleles differ from each other. The sequencing is not as pure as it was in Figure 3.9, as Um had a lower sequence quality. This figure shows the same as 3.9, that b1-3, b1-6 and b1-7 the best matches with our sequences.*

When we look at the same sequence area in Figure 3.11 that was shown in Figure 3.10, it is pretty clear that we again see all three of our sequenced products indicate β1-3, β1-6 and β1-7.

3.4.2.2 PCR targeting and dideoxy sequencing of Hbb2 PCR products

Hbb2, Hbb3 and Hbb4 are the globins that have the highest similarities among the cod globins. The exons of Hbb3 and 4 are highly similar. Most primer combinations for β2 are also primer combinations for β3 and β4 and were expected to give three PCR products (Table 3.7). This is because these three genes are so similar, that it can be difficult to make a suitable primer sequence that can separate them.

Table 3.7 *Primer combinations for Hbb2, 3 and 4 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
b234F-13	b234R417	868 gene β2 661 gene β3 658 gene β4	3000 1000 800 700 650
b234F-13	b234R447	411 gene β2 417 gene β3 413 gene β4	410
b234F255	b2R525	709	710
b234F255	b234R417	650 gene β2 415 gene β3 437 gene β4	420

b234F50	b234R447	346 gene β2 353 gene β3 349 gene β4	360

The primer combination of b234F255 and b2R525 was chosen for sequencing (Figure 3.12), as we here expected to detect only Hbb2. The gel electrophoresis showed one clean band with the expected size. The sequence had no errors, except for in the exact position where the β2-1 and β2-2 alleles differ from each other. cR has a G in the position where b2-2 has an A, but cR is also missing a base in the next position, which could be a sequencing error, but if you look at the electropherogram (Figure 3.13) you can see that in position 60 there is a double readout of a C and an A, which translates to G and T, which is the exact sequence b2-1 has and cR is therefore most likely b2-1.

```
b2-1            TCTCCCTCCCTCACACGTGCAGCTGCTGGCCGACTGCCTGACCGTCGTCATCGCCGCCAA 840
b2-2            TCTCCCTCCCTCACACGTGCAGCTGCTGGCCGACTGCCTGACCGTCGTCATCGCCGCCAA 840
cR-reverse      TCTCCCTCCCTCACACGTGCAGCTGCTGGCCGACTGCCTGACCGTCGTCATCGCCGCCAA 587
                ************************************************************

b2-1            GATGGGCACCAAATTCACCGTGGAGACCCAGGTGGCGTGGCAGAAGTTCCTGT-CTGTCG 899
b2-2            GATGGGCACCAAATTCACCGTGGAGACCCAGGTGGCATGGCAGAAGTTCCTGT-CTGTCG 899
cR-reverse      GATGGGCACCAAATTCACCGTGGAGACCCAGGTGGCG-GGCAGAAGTTCCTGTACTGTCG 646
                ********************************** *************** ******
```

Figure 3.12 Alignment of the sequences for PCR products from the primer pair b34F255 + b2R525. *b2R525 used as sequencing primer. Marked in purple is exon 3. Yellow is a difference between the two alleles. And teal is where the sequenced result differs from the published alleles.*

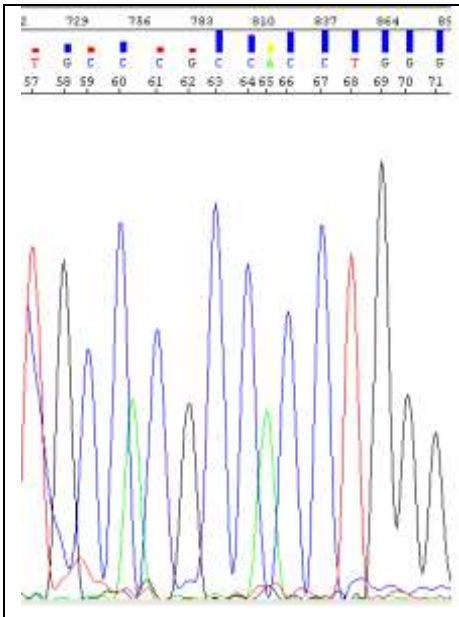

Figure 3.13 Electropherogram of cR. *This electropherogram shows how in position 60 there are two readouts, if they are counted as separate reads they give us CA which reverse complimentary is TG, which is exactly what b2-1 in figure 3.12 is in position 876-877.*

3.4.2.3 PCR targeting and dideoxy sequencing of Hbb3 and Hbb4 PCR products

β3 and β4 are very similar and our primers were not able to separate the two. Primers in Table 3.8 used in addition to primers from Table 3.6.

Table 3.8 *Primer combinations for Hbb3 and Hbb4 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
b34F-5	b234R447	408 gene β3 414 gene β4	410
b34F-5	b234R417	652 gene β3 646 gene β4	650

The sequence for β3 and β4 were close to that of the Genbank β3 and β4 (Figure 3.15), and the sequencing had a high quality on several of the runs. The reason for this is at β3 and β4 are very similar. fR seems at first glance seems to be β3, but if you look manually at the electropherogram (Figure 3.14) you can see that the sequence matches the known one from Genbank for both β3 and β4, but from the location where β3 and β4 differ from each other, the quality of the sequencing has dropped. If you set a manual reading frame for them, you can see that the two sequences are both there, just interlaced and interchanged, and you get the perfect sequence for both β3 and β4 as indicated in lower part of Figure 3.14.

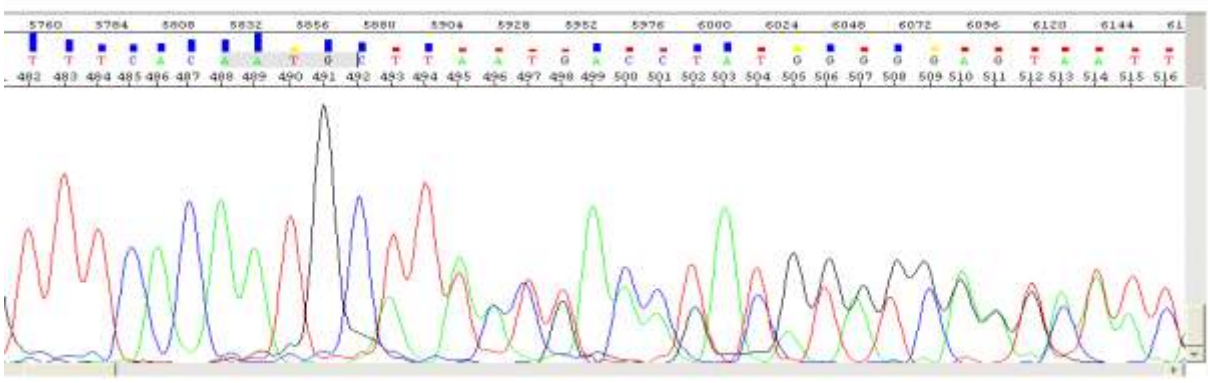

Figure 3.14 Electropherogram of sequence fR b34F-5 + b234R417. *The sequence has very high quality on the readout until the position where β3 and β4 differ from each other, where double peaks appear.*

The double readout of both β3 and β4 reads as shown below.

The red shows β3 and the teal shows β4. And it shows that our sequencing result is actually

both β3 and β4 intertwined.

```
β3:  AATTACTCCCCCATAGGTCATTAAGCATTGTGAAA
β4:  GTAGCTCGATATGTCTTTAGGAATGCATTGTGAAA
β3: -TTGCCCCATATGTCGTTAAGAATGCATTGTGAAA
β4: -TTGCCCCATATGTCGTTAAG----CATTGTGAAA
```

```
b4-1            TTGCCCCATATGTCGTTAAG----CATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 176
b4-3            TTGCCCCATATGTCGTTAAG----CATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 176
b4-2            TTGCCCCATATGTCGTTAAG----CATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 176
b4-4            TTGCCCCATATGTCGTTAAG----CATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 176
b3-1            TTGCCCCATATGTCGTTAAGAATGCATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 180
b3-3            TTGCCCCATATGTCGTTAAGAATGCATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 180
b3-2            TTGCCCCATATGTCGTTAAGAATGCATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 180
fR              CTCCCCCATAGGTCATTAAG----CATTGTGAAACTAAAACTCGTATTTATTTTTTTGAC 171
                 * ******* *** *****    ************************************
```

Figure 3.15 Alignment of the sequences for PCR products from the primer pair b34F-5 + b234R417. *b234R417 used as sequencing primer. Area marked in green is the β3 sequence from Figure 3.13*

3.4.2.4 PCR targeting and dideoxy sequencing of Hbb5 PCR products

β5 primers in general worked good

Table 3.9 *Primer combinations for Hbb5 and results of electrophoresis.*

Primer A	Primer B	Expected product	Found Product
b5F1	b5R449	657	650
b5F27	b5R449	621	650
b5F239	b5R709	482	480 460 290

The primer combination used for sequencing was b5F239 and b5R709 after a second PCR had been done to isolate the additional products found (Figure 3.16). The sequence here does not fit very well with the known sequences from Genbank. It is only about 120 nucleotides that have a high quality to for alignments. This is especially apparent from base 540 and onwards, but there had been a good line-up from base 346 and up to 540. All three 6L, 6m and 6s sequences appear to share the same differences in many positions. This indicates that there could be a pseudo gene, because there are too many differences to be a new allele. But poor sequencing quality is also a major possibility.

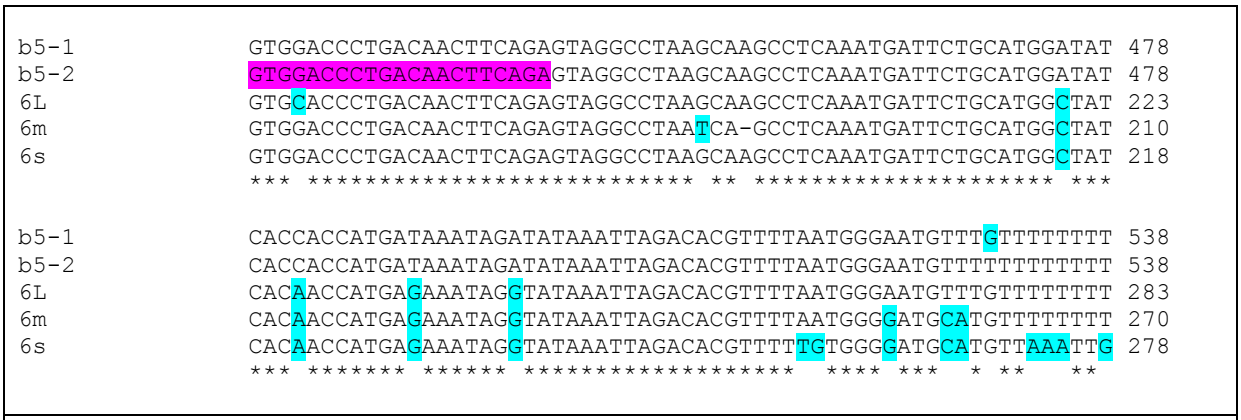

Figure 3.16 Primer combination of b5F239 + b5R709. *b5F239 used as sequencing primer. All three of our sequences share several of the same differences from the known alleles. The area marked in purple is exon 2 that goes from 217 – 439.*

3.5 454 Sequencing

3.5.1 Initial Testing of 454 Sequencing

The 454 sequencing primers are longer than the standard primers, and are 60-65 nucleotides long. They are in three sections, a gene specific part, a tag and a 454 specific part.

All the standard PCR primers were synthesized by DNA Technology, but TAG Copenhagen could offer cheaper primers for 454 sequencing.

Through other collaborators suspicions arose that the long primers, needed for 454 sequencing, provided by the new supplier where not of high enough quality. This was made clear during electrophoresis where they had low amounts of product and had smears on both sides of the product, and therefore also likely low quality product. New HPLC purified primers were ordered from the company and compared to the old ones, to rule out if there had been an error during production of these primers, or mishandled after they arrived. The purified primers showed somewhat less smearing, but did still not function well in the PCR experiments (Figure 3.17).

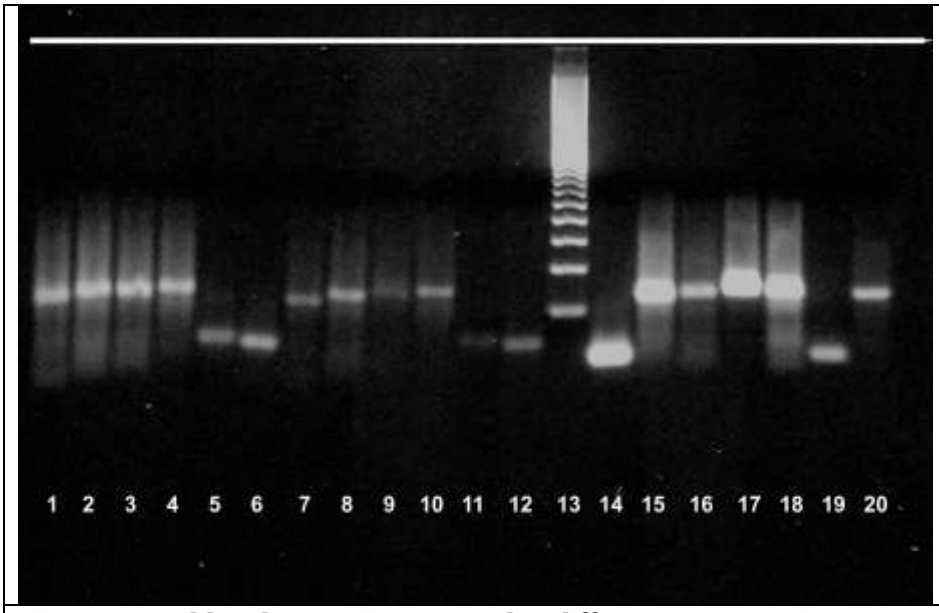

Figure 3.17 Old and new primers tested at different concentrations: ***New primers:*** *1: M1b234F-13 - 0.5 ul, 2: M1b234R417 - 0.5ul, 3: M1b5F1 – 0.5 ul, 4: M1b5R449 – 0.5 ul, 5: a4R676 – 0.5ul, 6: b2R614 – 0.5ul, 7: M1b234F-13 - 0.1 ul, 8: M1b234R417- 0.1 ul, 9: M1b5F1 – 0.1 ul, 10: M1b5R449 – 0.1 ul, 11: a4R676 – 0.1ul, 12: b2R614 – 0.1ul, 13: Molecular ruler 20bp,*
Old primers: *14: a3R1879 (1ul primer + 2ul water) 0.5 ul, 15: M1a1F14 – 0.5ul, 16: M1a1R400 – 0.5ul, 17: M1b34F-5 - 0.5ul, 18: M1b234R417 – 0.5 ul, 19: a3R1879 (1ul primer + 2ul water) 0.1ul, 20: M1a1F14 – 0.1ul*

Identical primers were ordered from DNA Technology, and tested against the original TAG Copenhagen, running a PCR experiment (Figure 3.18).

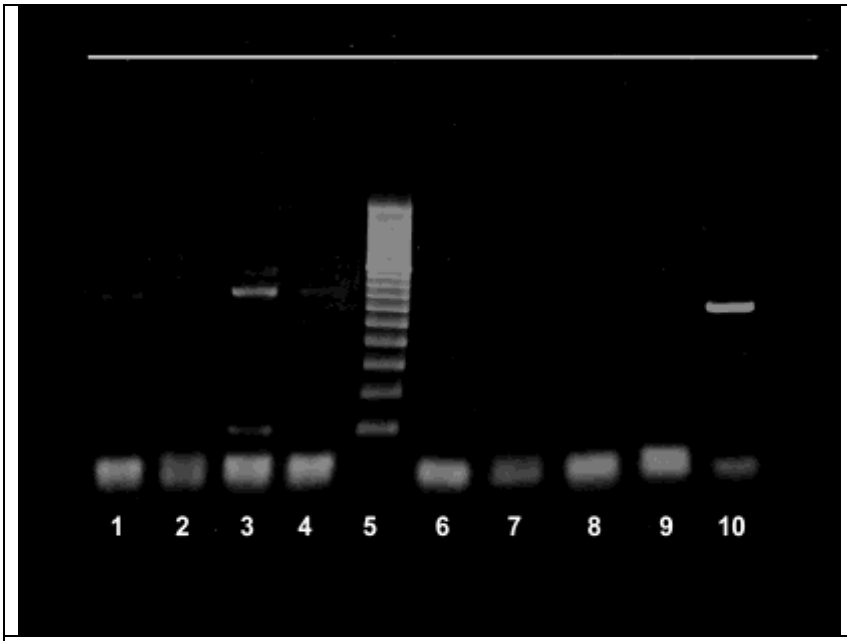

Figure 3.18 DNA Technology and TAG Copenhagen of primers tested against one other:
DNA Technology: *1: M1a3F59 and M1a3R733, 2: M1bF-26 and M1b1R611, 3: M1b234F-13 and M1b234R417, 4: M1b5F1 and M1b5R449, 5. Molecular ruler 100nt,*
TAG Copenhagen: *6: M1a3F59 and M1a3R733, 7: M1bF-26 and M1b1R611, 8: M1b234F-13 and M1b234R417, 9: M1b5F1 and M1b5R449, 10: Positive control.*
All primers from both companies were HPLC purified

Conclusion from the tests in Figure 3.18 is that the primers from DNA Technology gave PCR products, while TAG Copenhagen did not (Figure 3.18). Additionally when the primers themselves were run on agarose gels, the primers from DNA Technology were obviously more pure (Figure 3.19).

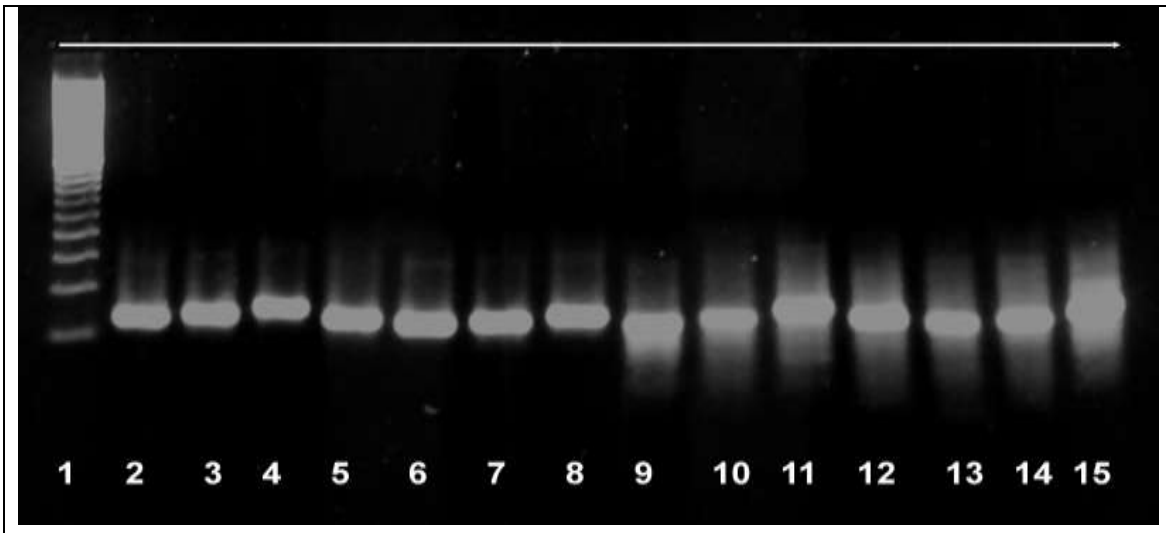

Figure 3.19 DNA Technology and TAG Copenhagen primers tested against one other.
DNA Technology primers in position 2 to 8, and TAG Copenhagen primers in position 9 to 15.
1: Molecular ruler 20nt, 2: M1a3F59, 3: M1a3R733, 4: M1b1F-26, 5: M1b1R611, 6: M1b234F-13, 7: M1b234R417, 8: M1b5F1, 9: M1a3F59, 10: M1a3R733, 11: M1b1F-26, 12: M1b1R611, 13: M1b234F-13, 14: M1b234R417, 15: M1b5F1
This image is overexposed to reveal the smears, bands were clearer when seeing the gel.

Figure 3.19 shows that the new primers have much less smear, and stronger bands.

4 Discussion:

4.1 PCR and agarose gels

The purification of PCR products from agarose gels, to use them as template for a second PCR, did not work well. It is possible that trace amounts of DNA has contaminated each of the separated bands of DNA, but more likely, it is that during the annealing step in the PCR, the primers have bound to pseudo genes as well as the intended targeted genes. This could lead to what was seen in electrophoreses, that we got several bands from a single primer pair, even when all alleles should produce a similar product. This would also explain why purification from electrophoreses did not successfully purify the products, because if the pseudo genes and intended genes are in the product, it is possible that both these genes will bind together as a hybrid of the normal gene and pseudogene, thereby creating the third of the bands we saw with many primer pairs, where two of the bands would be double stranded intended gene and double stranded pseudogene, and the third one would be a single strand of both.

Another more likely explanation for the multiple bands on some gels runs, is that some genes are so similar, that during annealing in PCR the single strands of DNA can bind to similar, but of a different gene sequences, thus resulting in a hybrid of the two different genes (Figure 4.1 and Figure 4.2).

Figure 4.1 Illustration of β234 DNA strands.
The blue bars indicate the primers. The black line is the shared gene sequence of β2, β3 and β4, and the red line is the difference between β2 and β3/β4.

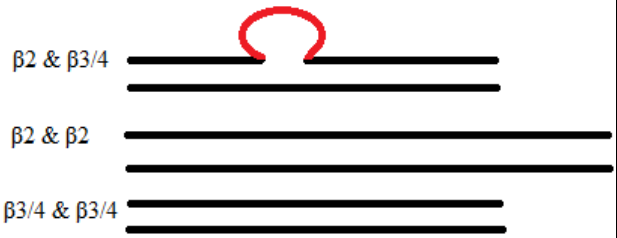

Figure 4.2 Illustration of hybrid DNA strands of β234.
This illustration shows how genes that are very similar can make hybrid pairs of DNA strands during annealing of PCR. If these were run in a gel the hybrid of β2&β3/4 would show an unknown band in the gel, probably in-between the expected result for β2 and β3/4.

The quality of the first batch of the long primers needed for 454 sequencing was also an issue. These had smears and low quality bands, this shows how vigilant one has to be in every single step of a project, as there can be errors from one self as well as a supplier of tools and supplies.

The primer design in general has worked well, especially for α primers. They gave very clear clean results. This was expected because of the difference between the α genes is bigger than between some of the β genes. Some of the β primers need to be looked at further to see why they did not work. The primers were though able to separate the β234 complex, and with manual interpretation of sequencing through electropherogram it was possible to see both β3 and β4 sequence, this is something I had not expected to be able to see before hand, because β3 and β4 are so similar as shown in Table 1.2.

4.2 Sequencing

Several results indicate that there could be unknown alleles or pseudogenes. A example of a possible pseudogene is Figure 3.16 where all three of our products show the same deviations in several positions from the two known alleles of β5, seeing as all but one of the deviations of β5 in our results are in introns make it more likely that this is a new allele, because the exons are much more preserved than the introns are. The indication that our β5 results are a pseudogene and not a new allele, is most apparent when comparing both known alleles of β5, β5 is 725 nucleotides long for both alleles and there is only a single nucleotide difference between the two, this is on nucleotide 530 which is in an intron (Figure 4.3), so finding a new allele with multiple differences is possibly not likely. But further testing is needed to determine if this is the case.

```
b5-1            CCACCATGATAAATAGATATAAATTAGACACGTTTTAATGGGAATGTTTGTTTTTTTTCG 540
b5-2            CCACCATGATAAATAGATATAAATTAGACACGTTTTAATGGGAATGTTTTTTTTTTTCG 540
                ********************************************** **********
```

Figure 4.3 β5-1 and β5-2 compared to each other. Both known β5 alleles compared to each other and highlighted the only variation between the two.

4.3 Future Projects

A continuation of this project would be to setup and test 454 sequencing with our long primers. This is a much faster method of sequencing as you can do up to one million at a time.

Other projects could go research with genetic if there are other sub populations of cod on Faroe Bank or the Faroe plateau, and go further in-depth by looking for genetic differences between Faroese cod.

Future projects could focus on some of the possible new alleles. Where primers could be designed specifically to target these, and figure out if these are new unknown alleles or pseudogenes. Should new alleles be discovered, other projects could then focus on comparing the different alleles of the genes, to see if there is any positive or negative results between them for growth patterns of the cod. Further projects again could find out if any of the variations are show different results to varying amounts of oxygen and water temperatures. This could be part of research of how global warming affects the gene expressions of haemoglobin in cod.

Most fishing for cod around the Faroe Islands is done with various kinds of nets, these introduce a selection pressure where it is fittest to be smaller than the meshes of the nets. A future project could be to find out if or how this selection pressure affects haemoglobins expressed in cod.

Another project that would be interesting is to research how old the Faroese cod populations are, and if there's how the genetic drift for all north Atlantic cod populations compares to each other. And if there is any difference in the genetic drift, then to compare the environments and try to find out why.

Knowing the alleles present is important for population genetic studies and stock management in general.

5 Conclusion

The primers for α genes worked very well, and the primers for β genes were less selective as expected since several of the β genes are so similar. Testing of the long primers can continue, and seem to be able to get results.

Though more testing is needed to draw any real conclusion on whether there are any new alleles in my testing results, or if the observed variations are pseudogenes or hybrids of two different DNA strands.

6 References

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raaf, M., Roberts, K. and Walter, P. 2009. Essential Cell Biology - third edition. Garland Science, New York.

Antonini, E., Brunori, M. 1971. Hemoglobin and Myoglobin in their Reactions with Ligands. North-Holland Pub. Co.

Borza, T., Stone, C., Gamperl, A. K. and Bowman, S. 2009. Atlantic Cod (*Gadus morhua*) hemoglobin genes: multiplicity and polymorphism. BMC Genetics, 10: 51.

Brix, O., Thorkildsen, S. and Colosimo, A. 2004. Temperature acclimation modulates the oxygen binding properties of the Atlantic cod (*Gadus morhua L.*) genotypes—HbI*1/1, HbI*1/2, and HbI*2/2—by changing the concentrations of their major hemoglobin components (results from growth studies at different temperatures). Comparative Biochemistry and Physiology, Part A 138: 241

Coyle, T. 1997. Stock identification and fisheries management: the importance of using several methods in a stock identification study. Austrilian Society for Fish Biology.

Frank, K. T., Petrie, B., Fisher, J. A. D. and Leggett, W. C. 2011. Transient dynamics of an altered large marine ecosystem. Nature 477: 86

Fry, F. E. J. and Hochachka, P. W. 1970. Comparative physiology of thermoregulation vol. 1 (ed. Whittow, G. C.). Academic, New York.

Hardison, R. 1999. The evolution of hemoglobin: Studies of a very ancient protein suggest that changes in gene regulation are an important part of the evolutionary story. American Scientist 87: 126

Hartl, D. L. and Jones, E. W. 2006. Essential Genetics A Genomic Perspective - Fourth Edition. Jones and Bartlett Publishers. Sudbury.

Hartl, D. L. and Clark, A.G. 2007. Principles of Population Genetics - Fourth Edition. Sinauer Associates, Inc. Publishers. Sunderland.

Hoffmann, F. G., Opazo, J. C., Hoogewijs, D., Hankeln, T., Ebner, B., Vinogradov, S. N., Bailly, X. and Storz, J. F. 2012. Evolution of the Globin Gene Family in Deuterostomes: Lineage-Specific Patterns of Diversification and Attrition. Mol. Biol. Evol. Advance Access.

Magnussen, E. 1993. Why is the cod growing faster on the Faroe Bank than on the Faroe Plateau? Growth experiments and genetic investigations of cod from the two areas (in Danish) [*Hvorfor vokser torsken hurtigere på Færøbanken end på Færøplateauet? Vækstforsøg og genetiske undersøgelser af torsk fra de to områder*]. MSc thesis, University of Copenhagen. 1-104, 2003.

Magnussen, E. 2007. Interpopulation comparison of growth patterns of 14 fish species on Faroe Bank: are all fishes on the bank fast-growing? Journal of Fish Biology 71, 453

Magnussen, E. 2009. Tá bankatoskurin hvarv í Tangafirði (Frøði 2/2009)

Märss, T. 2001. Andreolepis (*Actinopterygii*) in the Upper Silurian of Northern Eurasia. Proc. Estonian Acad. Sci. Geol. 50: 174

Petersen, M. F. and Steffensen, J. F. 2002. Preferred temperature of juvenile Atlantic cod (*Gadus morhua*) with different haemoglobin genotypes at normoxia and moderate hypoxia. Journal of Experimental Biology 206: 359-364

Shelton, P. A. 2005. Did over-reliance on commercial catch rate data precipitate the collapse of northern cod? ICES Journal of Marine Science, 62: 1139.

Sick, K. 1961. Haemoglobin polymorphism in fishes. Nature 192: 894.

Sick, K. 1965. Haemoglobin polymorphism of cod in the North Sea and the North Atlantic Ocean. Hereditas 54: 49.

Sundby, S. 2000. Recruitment of Atlantic cod stocks in relation to temperature and advection of copepod populations. Sarsia 85: 277

Wetten, F. O., Nederbragt, A. J., Wilson, R. C., Jakobsen, K. S., Edvardsen, R. B. and Andersen, Ø. 2010. Genomic organization and gene expression of the multiple globins in Atlantic cod: conservation of globin-flanking genes in chordates infers the origin of the vertebrate globin clusters. BMC Evolutionary Biology 10: 315

7 Acknowledgements

First I would like to thank the Faculty of Science and Technology at the University of the Faroe Islands (NVD) for use of equipment and facilities.
I would also like to thank Marjun Mortensen (laboratory technician at NVD) for her guidance and time in the laboratory and Debes H. Christiansen (Faroese Food- and veterinary authority) for the sequencing work.
Finally a special thank you to my two supervisors Svein-Ole Mikalsen (Professor in molecular biology at NVD) and Eyðfinn Magnussen (Associate professor of biology at NVD) for their guidance and patience.

8 Appendix

Appendix A – Primer combinations

gDNA						
Primer F	Primer Sequence	Primer R	Primer Sequence	Temperature	Expected Product	Found Product
a1F14	CAAAGGACAAGGCGACCGTCA	a1R400	AAGCACACAGGAACTTGTCGAC (GTCGACAAGTTCCTGTGTGCTT)	60	599	~570
a1F14	CAAAGGACAAGGCGACCGTCA	a1R414	AGCTCGCTCAGAGTGAGAAGAC (GTCTTCTCACTCTGAGCGAGCT)	60	356	~330
a2F-27	CTAGAAAGCAACTATCTGAACGTCAAC	a2R346	GGAAGATCATGGACATGCACACCA (TGGTGTGCATGTCCATGATCTTCC)	60	592	~540
a2F312	CTCTCACTGGAAGGACGCGAG	a2R346	GGAAGATCATGGACATGCACACCA	60	308	~290
a2F352	AGGAAGCATGGCATCACCATCATG	a2R346	GGAAGATCATGGACATGCACACCA	60	268	~270
a3F59	TGGCTGACCGCATAGGAGCAG	a3R394	ACGCAGAGAGATACTTGTCCATG (CATGGACAAGTATCTCTCTGCGT)	60	1701	1600
a3F851	ACCTCGCTCCTCTTCAAACCTAC	a3R394	ACGCAGAGAGATACTTGTCCATG		953	1000
a3F851	ACCTCGCTCCTCTTCAAACCTAC	a3R1304	CCTATACCTATCTGTCTAGTACAC (GTGTACTAGACAGATAGGTATAGG)	60	453	1200 400
a3F1163	GGTTAACTAGGACCATAGGTGTAC	a3R394	ACGCAGAGAGATACTTGTCCATG	60	640	600
a4F63	AGCAGTGGAGATCGGACACCAG	a4R614	CTGAGGAGTGAAATCATCTGGGAAC (GTTCCCAGATGATTTCACTCCTCAG)	60	543	460
a4F249	CCAAACCAAAGTCTACTTCTCCCA	a4R614	CTGAGGAGTGAAATCATCTGGGAAC	60	365	320
b1F-26	CATTGAACCCTTAAGACWACGCCAC	b1R252	GGACTTGATTTCGTCCATGTGGTC (GACCACATGGACGAAATCAAGTCC)	59 60 (16/3)	544	(16/3) 500 smudge 530 clear
b1F-26	CATTGAACCCTTAAGACWACGCCAC	b1R463b	TAGCGTCAGCACACATCTTCTAGTG (CACTAGAAGATGTGTGCTGACGCTA)	59	1182	blank
b1F-26	CATTGAACCCTTAAGACWACGCCAC	b1R463	TAGCGTCAGCACACATCTTCTAG (CTAGAAGATGTGTGCTGACGCTA)	59	1182	blank
b1F-26	CATTGAACCCTTAAGACWACGCCAC	b1R611	CACTCCGAGAAACTGCACGTCGA	60	584	580 600 630
b1F-26	CATTGAACCCTTAAGACWACGCCAC	b1R261	GACGAAATCAAGTCCACCTACGCT	60	530	380

						400 420
b1F-44	GAGATTCAGCCTAAGCTACATTGAAC	b1R252	GGACTTGATTTCGTCCATGTGGTC	59 60 (16/3)	562	Blank 540 (16/3)
b1F-44	GAGATTCAGCCTAAGCTACATTGAAC	b1R463b	TAGCGTCAGCACACATCTTCTAGTG	59	1200	Blank
b1F-44	GAGATTCAGCCTAAGCTACATTGAAC	b1R463	TAGCGTCAGCACACATCTTCTAG	59	1200	Blank
b1F-44	GAGATTCAGCCTAAGCTACATTGAAC	b1R611	CACTCCGAGAAACTGCACGTCGA	60	609	220 580 600 630
b1F-44	GAGATTCAGCCTAAGCTACATTGAAC	b1R261	AGCGTAGGTGGACTTGATTTCGTC (GACGAAATCAAGTCCACCTACGCT)	60	571	220 380 400 420
b1F44	TCTGGAGCAAGATCGACATTGATG	b1R252	GGACTTGATTTCGTCCATGTGGTC	59	475	Blank
b1F44	TCTGGAGCAAGATCGACATTGATG	b1R463b	TAGCGTCAGCACACATCTTCTAGTG	59	1113	Blank
b1F44	TCTGGAGCAAGATCGACATTGATG	b1R463	TAGCGTCAGCACACATCTTCTAG	59	1113	Blank
b1F44	TCTGGAGCAAGATCGACATTGATG	b1R611	CACTCCGAGAAACTGCACGTCGA	60	522	Blank
b1F44	TCTGGAGCAAGATCGACATTGATG	b1R261	GACGAAATCAAGTCCACCTACGCT	60	485	Blank
b1F589	CACTCCGAGAAACTGCACGTCGA	b1R463b	TAGCGTCAGCACACATCTTCTAGTG (CACTAGAAGATGTGTGCTGACGCTA)	59	614	ca 600 (16/3)
b1F589	CACTCCGAGAAACTGCACGTCGA	b1R463	TAGCGTCAGCACACATCTTCTAG	59	614	ca 650 (16/3)
b1F-35	CCTAAGCTACATTGAACCCTTAAGAC	b1R463b	TAGCGTCAGCACACATCTTCTAGTG	60	1191	
b1F-35	CCTAAGCTACATTGAACCCTTAAGAC	b1R463	TAGCGTCAGCACACATCTTCTAG	60	1191	Blank
b1F-35	CCTAAGCTACATTGAACCCTTAAGAC	b1R611	TCGACGTGCAGTTTCTCGGAGTG (CACTCCGAGAAACTGCACGTCGA)	60	600	580 600 630
b1F-35	CCTAAGCTACATTGAACCCTTAAGAC	b1R252	GGACTTGATTTCGTCCATGTGGTC	60	552	530 (16/3)
b1F-35	CCTAAGCTACATTGAACCCTTAAGAC	b1R261	GACGAAATCAAGTCCACCTACGCT	60	563	380 400 420
b234F-13	CACAWCAGCAACCATGGTTGAG	b234R417	GACGACGACAGACAGGAACTTCTG (CAGAAGTTCCTGTCTGTCGTCGTC)	59	868 allel 2 661 allel 3 658 allel 4	3000 1000 800 700

						650
b234F-13	CACAWCAGCAACCATGGTTGAG	b234R447	CTTGTCAGAGTGGAGCAGAGAC (GTCTCTGCTCCACTCTGACAAG)	59	411 allel 2 417 allel 3 413 allel 4	410
b234F255	TCTGATCGTGTACCCCTGGACC	b2R525	AGATCACCCACTCTGCGGCTTG (CAAGCCGCAGAGTGGGTGATCT)	59	709	710
b234F255	TCTGATCGTGTACCCCTGGACC	b234R417	GACGACGACAGACAGGAACTTCTG	59	650 allel 2 415 allel 3 437 allel 4	420 with trace after BL
b234F50	CAMCTTGGACTACGAAGAGATCGG	b234R447	CTTGTCAGAGTGGAGCAGAGAC	59	346 allel 2 353 allel 3 349 allel 4	360
b34F-5	CACCATGGTTGAGTGGACAGATARTG	b234R447	CTTGTCAGAGTGGAGCAGAGAC	59	408 allel 3 414 allel 4	410
b34F-5	CACCATGGTTGAGTGGACAGATARTG	b234R417	GACGACGACAGACAGGAACTTCTG	59	652 allel 3 646 allel 4	650
b5F1	ATGGTAGAGTGGACAGAATTTGAG	b5R449	CGCTACTAGTGGTACTGGCTTC (GAAGCCAGTACCACTAGTAGCG)	59	657	650
b5F27	CGATACAATCAAGGACATCTTCTC	b5R449	CGCTACTAGTGGTACTGGCTTC	59	621	650
b5F239	ACTCGGAGGTACTTTGGAAACTTC	b5R709	ACGGTAGTCCTACTGTCCATGCA (TGCATGGACAGTAGGACTACCGT)	59	482	1000 460 290 weak 27/3 480 27/3