ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

BIOLOGIAKO GRADUA/
GRADO EN BIOLOGÍA/
DEGREE IN BIOLOGY

GRADU AMAIERAKO LANA/
TRABAJO DE FIN DE GRADO/
BACHELOR´S THESIS

ANNOTATION OF ORTHOLOGOUS GENES IN *THUNNUS THYNNUS* AND *T. ALBACORE* SPECIES.

ASIER OLALDE BELTRAN DE HEREDIA

Leioa, 2014ko uztaila/ Julio 2014/ July 2014

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

INDEX

ABSTRACT

Albacore and Atlantic Bluefin tuna are two pelagic fish. Atlantic Bluefin tuna is included in the IUCN red list of threatened species and albacore is considered to be near threatened, so conservation plans are needed. However, no genomic resources are available for any of them. In this study, to better understand their transcriptome we functionally annotated orthologous genes. In all, 159 SNPs distributed in 120 contigs of the muscle transcriptome were analyzed. Genes were predicted for 98 contigs (81.2%) using the bioinformatics tool BLAST. In addition, another bioinformatics tool, BLAST2GO was used in order to achieve GO terms for the genes, in which 41 sequences were given a biological process, and 39 sequences were given a molecular process. The most repeated biological process was metabolism and it is important that no cellular process was given in any of the sequences. The most abundant molecular process was binding and very few catalytic activity processes were given. From the initial 159 SNPs, 40 were aligned with a sequence in the database after BLAST2GO was run, and were polymorphic in Atlantic Bluefin tuna and monomorphic in albacore. From these 40 SNPs, 24 were located in an open reading frame of which four were non-synonymous and 20 were synonymous and 16 were not located in a known open reading frame,. This study provides information for better understanding the ecology and evolution of these species and this is important in order to establish a proper conservation plan and an appropriate management.

Hegalaburra eta hegaluzea, bi arrain pelagiko dira. IUCN-ren espezie mehatxatuen zerrenda gorrian dago hegalaburra eta hegaluzea ia mehatxatuta dago eta ondorioz, kontserbazio planen beharra dago. Hala ere, ez dago ez bata ez bestearen inolako baliabide genetikorik. Ikerketa honetan, transkriptoma hobeto ulertzeko helburuarekin, gene ortologoen anotazio funtzionala egin da. Guztira, muskuluko transkriptoma-ko 120 kontigetan banatutako 159 SNP aztertu ziren. BLAST tresna bioinformatikoa erabiliz, geneak iragarri ziren 98 (%81,2) kontigetarako. Horrez gain, BLAST2GO deritzon beste tresna bioinformatiko bat erabili zen geneei dagozkien GO terminoak lortzeko. 41 sekuentzietarako prozesu biologiko jakinak lortu ziren eta prozesu molekularrak 39 sekuentzietarako lortu ziren. Prozesu biologikoei dagokionez, ugariena metabolismoa izan zen, eta aipatzekoa da ere prozesu zelularrik agertu ez izana. Prozesu molekularrei dagokienez, lotura izan zen ugariena eta prozesu katalitikoen agerpena urria izan zen. BLAST2GO egin eta gero, hasierako 159 SNP-etatik, 40 alineatu egin ziren datu baseko sekuentziaren batekin eta gainera, hegalaburrean polimorfikoak eta hegalaburrean monomorfikoak ziren. 40 SNP hauetatik, 24 “open reading frame” batean zeuden kokatuta eta hauetatik 4 ez-sinonimoak eta 20 sinonimoak ziren eta beste 16ak ez zeuden kokatuta “open reading frame” batean. Ikerketa honek, ematen du espezie honen ekologia eta eboluzioa hobeto ulertzeko informazioa eta hau, garrantzitsua da kontserbazio plan eta ustiaketa egoki bat ezarri ahal izateko.

INTRODUCTION

Large-scale genome sequencing has resulted in more evolutionary and functional analyses in the field of comparative genomics. Conserved sequences can be used for the reconstruction of the evolution of species as well as for genomic annotation. The concept of orthology comes from molecular systematics (Fitch, 1970). Lately, orthology has been used for functional characterization and classification in the comparison of genomes (Chervitz et al., 1998; Mushegian et al., 1998; Rubin et al., 2000; Tatusov et al., 1997; Tatusov et al., 2000; Wheelan et al., 1999). Orthologous genes are those that evolved from a common ancestor by speciation and it is likely that their function is conserved overtime and thus, analyses of these genes can be used for gene annotation. The clustering of orthologs allows highlighting the divergence and conservation of gene families and biological processes (Fitch, 1970).

This study focuses on orthologous genes in two widely distributed pelagic tunas: Atlantic Bluefin tuna (*Thunnus thynnus* L., 1758) and albacore (*Thunnus alalunga* Bonn, 1788). Atlantic Bluefin tuna is one of the largest species in the genus *Thunnus*, whereas albacore is one of the smallest tunas. Nowadays, Atlantic Bluefin tuna is spread through the North Atlantic Ocean and the Mediterranean Sea (Figure 1). An analysis of present over historical ranges showed that this species has had larger range contractions (minus 46% since 1960) than any other pelagic species (Worm and Tittensor, 2011). It has been estimated that this species has declined at least 51% over the past three generation lengths (39 years) and it is listed as Endangered under Criterion A2 (Collette et al., 2011a). This has resulted in a decrease of this species to historical levels due to massive overfishing (Fromentin and Powers, 2005; MacKenzie et al., 2009; Majkowski, 2007). Although there is an important trans-Atlantic migration of individuals, there are at least three reproductively isolated stocks: The western Atlantic stock, the eastern Atlantic stock and the Mediterranean Sea. Albacore has a wider distribution as it is cosmopolitan in tropical and temperate waters of all the oceans including the Mediterranean Sea, but it is not present at the surface between 10°N and 10°S (Collette, 2001) (Figure 2). This species is important for many commercial fisheries around the world and there are six stocks that are globally managed. Since 2004, the North Atlantic stock is overexploited, the Indian Ocean and North Pacific are Fully Exploited, the South Atlantic and South Pacific are Moderately Exploited, and the Mediterranean is unknown (Majkowski, 2007). This species is considered to be near threatened (Collette et al., 2011b). This means that there is a need of conservation and population management efforts.

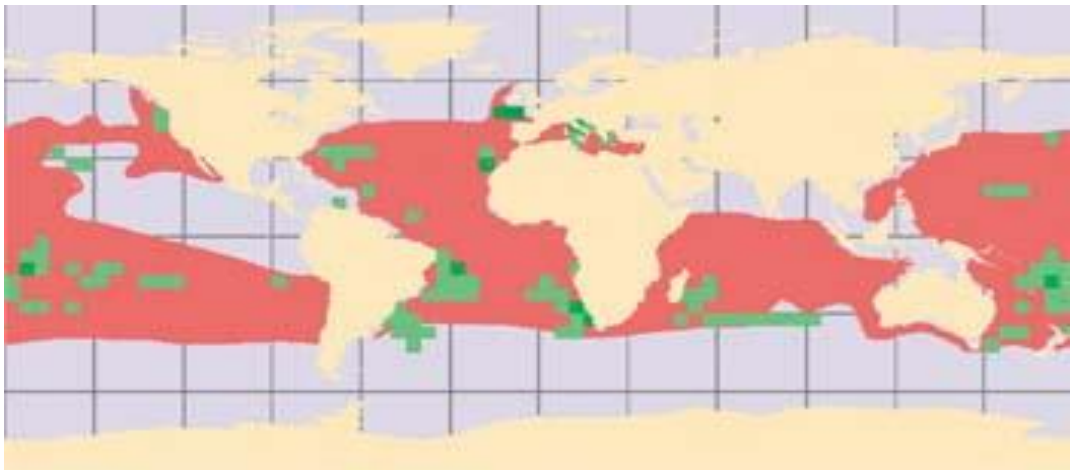

Figure 1. Global distribution of *Thunnus alalunga* (in red) and main fisheries areas (green) based on captures during the period 2000-2005 (Maguire, 2006).

Figure 2. Global distribution of *Thunnnus thynnus* (in red) and main fisheries areas (green) based on captures during the period 2000-2005 (Maguire, 2006).

From 1980 on, many papers dealing with conservation and genetics have focused on the description and identification of individuals, genetic population structure, kin relationships, and taxonomic relationships (Allendorf el al., 2013). It is thought that genomics will gain importance in the field of conservation biology (Avise, 2010; Frankham, 2010; Ouborg et al., 2010; Primmer, 2009). The explosion of information due to the available complete genome sequences are transforming how we understand the amount, distribution and functional significance of genetic variation in natural populations (Allendorf et al., 2010; Allendorf et al., 2013; Amato et al., 2008).

Nowadays, there is no reference genome of any of the species within the genus *Thunnus* and thus, there is little that is known genetically about this genus. Nevertheless, our research group is part of a consortium, whose goals include obtaining the transcriptome and genome of the Atlantic Bluefin tuna. Sequences obtained from the transcriptome of the Atlantic Bluefin tuna were used to identify SNPs and genotype them in albacore (Laconcha et al., *unpublished*). This way, 240 sequences were identified in which SNPs were located and these SNPs were genotyped cross-species. As these SNPs were amplified in two different species, they should be located in highly conserved genes in these two species.

Objectives

The main goal of this study is to do a gene annotation in order to predict the function of some conserved orthologous genes in Atlantic Bluefin tuna and albacore based on some contigs from the muscle trasncriptome of the Atlantic Bluefin tuna. Gene Ontology (GO) annotation was made to predict the function of those genes. Also, SNPs within those contigs that are polymorphic in the Atlantic Bluefin tuna and monomorphic in albacore have been analyzed to infer if any nucleotide change has resulted in an amino acid change as this could be underlying an adaptation (DeWoody et al., 2010).

MATERIALS AND METHODS

The analyzed sequences were taken from a project named "Biological and genetic sampling and analysis" of the ICCAT Atlantic-Wide Research Program on Bluefin Tuna. In this project, a massive sequencing of the muscle transcriptome of the Atlantic Bluefin Tuna was carried out by using the next-generation sequencer Roche 454 Genome Sequencer. There were 240 contigs in the project. From those, 120 contigs were taken for the present study as the other 120 contigs were taken for another study and that is enough in order to do this study and achieve the goals. There were 159 SNPs within the 120 contigs. All these SNPs are polymorphic in the Atlantic Bluefin Tuna but some of them are monomorphic in albacore.

In this study, gene annotation and GO terms of the transcriptome were analyzed. In addition, those SNPs that were located in protein coding sequences were analyzed to observe if there was any amino acid change.

Gene prediction through alignment

First, in order to predict genes, BLAST (Basic Local Alignment Search Tool) was used. BLAST is usually used to input a nucleotide or protein sequence as a query against all or a subset of the public sequence databases, pasting the sequence into the textbox on one of the BLAST webpages. This bioinformatics tool is the most frequently used one for calculating sequence similarity (Fassler and Cooper, 2011; Altschul et al., 1990). Sequences must be introduced in FASTA format. The 120 contigs were introduced into BLAST and a nucleotide BLAST (Altschul et al., 1990; Wheeler et al., 2007) was run (Figure 3). In a nucleotide BLAST, a nucleotide query is used and the search is done in the nucleotide database. In this study, the Nucleotide collection (nr/nt) database was used and the search was done by using the "teleost fishes" database. Finally, the blastn (somewhat similar sequences) BLAST algorithm was used, as this is the most adequate for finding alignments to relate nucleotide sequences from other organisms. Only those hits in which E value was lower than 10^{-5} were considered. The Expectation value or Expect value (E value) represents the number of different alignments with scores equivalent or better than S that is expected to occur in a database search by chance. The lower the E value, the more significant the score and the alignment (Madden, 2002). By doing this, the genes within the contigs were predicted.

Gene Ontology (GO)

Then, the SNPs within those contigs that were aligned with a sequence were considered for further analysis. After this, the contigs were introduced into ORF Finder to see which of the SNPs were located in known open reading frames. ORF Finder (Open Reading Frame Finder) is a graphical analysis tool, which finds all open reading frames of a selectable minimum size in a user´s sequence or in a sequence already in the database. This bioinformatics tool identifies all open reading frames using the standard or alternative genetic codes. In this study, the standard genetic code was used. ORF Finder gives six different possible results so in order to be considered for the following steps the SNPs had to be in an open reading frame at least in one of those (Wheeler et al., 2007). Then, the contigs were shortened. They were shortened to take a shorter sequence that included the SNP that was to be analyzed. This was done because contigs were too long compared to mean exon length, which in humans for example, it is about 122bp with little length variation (http://info.gersteinlab.org/Genome_Statis_tics). Afterwards, these shortened sequences were introduced into Blast2GO where a BLASTX and a GO annotation were run. Blast2GO is a research tool designed with the main purpose of enabling Gene Ontology (GO) based data mining on sequence data for which no GO annotation is yet available. Blast2GO joints one application

GO annotation based on similarity searches with statistical analysis. This free tool offers a suitable platform for functional genomics research in non-model species. Blas2GO is an intuitive and interactive desktop application that allows monitoring and comprehension of the whole annotation and analysis process (Conesa et al., 2005). The BLASTX searches protein database using a translated nucleotide query. Blast2GO uses BLAST to find homologs to fasta formatted input sequences (Conesa et al., 2005). Those sequences that showed a significant hit (E value < 10^{-6}) and had the SNP within them aligned were considered. This was done in order to obtain the predicted gene, the reading frame and the E value of the BLASTX hit as well as molecular and biological processes for those sequences. Then, biological processes were gathered in larger groups based on their function. For this, the information found in The GO consortium was used (http://geneontology.org/GO.contents.doc. shtml#ontology). Finally, among these sequences just those that were polymorphic in Bluefin Tuna and monomorphic in albacore were inserted in ORF finder in order to see which of the SNPs were protein coding for annotated genes and among those which ones were synonymous and non-synonymous (Figure 3).

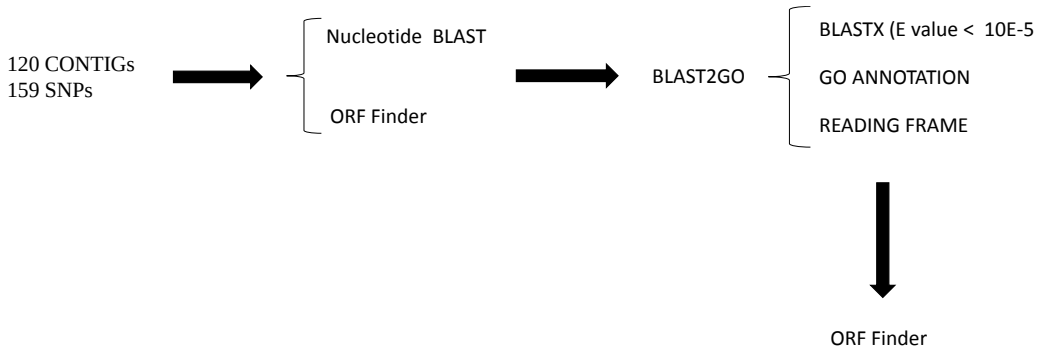

Figure 3: The steps of the analysis of the contigs and SNPs.

RESULTS

Gene prediction

After the BLASTN was run, from the starting 159 SNPs within the 120 contigs, 133 SNPs within 98 contigs were aligned and had a hit and thus, a gene was predicted for each of the contigs (annexes table 1).

GO annotation

From these 133 SNPs, 99 SNPS appeared to be in an open reading frame.

The contigs were then shortened into 89 sequences including those 99 SNPs.

After running the BLASTX in the Blast2GO program the 56 shortened sequences matched with a known sequence from the NCBI database and the matches had an E value lower than 10^{-6} (annexes table 2).

By running the Blast2GO, the GO terms were given for each of the 56 sequences (annexes table 3). On the one hand, there are biological processes. The number of sequences in which biological process appear is shown in Table 1. The most common biological processes were the regulation of biological process, catabolic processes, multicellular organismal development, translation and cellular component organization. These five biological processes comprise nearly the 40% (38.11%) of all the annotations and are just the 13.89% of all biological processes shown in Table 1. The number of biological processes is higher than that of the sequences (56) because most of the contigs received more than one biological process (Table 1).

Table 1. Number of sequences (N) from those introduced into Blast2GO that have been annotated with each of the Biological processes along with the percentage.

Biological process	N	Percentage
Regulation of biological process	12	8,63
Catabolic process	12	8,63
Multicellular organismal development	10	7,19
Translation	10	7,19
Cellular component organization	9	6,47
Nucleobase containing compound metabolic process	7	5,04
Generation of precursor metabolites and energy	7	5,04
Biosynthetic process	7	5,04
Cell differentiation	6	4,32
Carbohydrate metabolic process	5	3,60
Metabolic process	5	3,60
Anantomical structure morphogenesis	4	2,88
Cytoeskeleton organization	4	2,88
Response to stress	4	2,88
Reproduction	4	2,88
Transport	4	2,88
Ion transport	3	2,16

Biological process	N	Percentage
Cell death	3	2,16
Response to abiotic stimulus	2	1,44
Protein metabolic process	2	1,44
Signal transduction	2	1,44
Viral process	2	1,44
Cell cycle	2	1,44
Cellular protein modification process	1	0,72
Cell recognition nucleotide binding	1	0,72
Actin cytoeskeleton organization	1	0,72
Response to external stimulus	1	0,72
Response to biotic stimulus	1	0,72
Structural molecule activity	1	0,72
Single-organism transport	1	0,72
DNA metabolic process	1	0,72
Embryo development	1	0,72
Cellular homeostasis	1	0,72
Protein transport	1	0,72
Protein binding	1	0,72
Growth	1	0,72
Total	139	100

From the total 139 biological processes the most abundant ones were those related to metabolic process (28%) and developmental process (34%) (Table 2). The second most abundant biological process was regulation (9.35%).

Table 2. Biological processes groups that resulted when the biological processes were grouped into larger units based on the similarities of the processes. (N: number of biological processes in the group; %: percentages).

Biological processes groups	N	%
Metabolic process	39	28,06
Developmental process	34	24,46
Regulation	13	9,35
Response to stimulus	8	5,76
Transport	8	5,76
Others	37	26,62
Total	139	100,00

On the other hand, there are molecular processes. In this case, there were four molecular processes: nucleotide binding, structural molecule activity, actin binding, binding, which comprised 39% of all the molecular processes and were just the 17.39% of the all the molecular processes shown (Table 3). The molecular processes were not equally distributed, as some are more abundant (Nucleotide binding appeared 10 times) whereas others appeared just once (Lipid binding).

Table 3. Number of sequences (N) from those introduced into Blast2GO that have been annotated with each of the Molecular processes along with the percentage.

Molecular process	**N**	Percentage
Nucleotide binding	10	12,20
Structural molecule activity	8	9,76
Actin binding	7	8,54
Binding	7	8,54
Catalytic activity	6	7,32
Protein binding	6	7,32
Calcium ion binding	5	6,10
Hydrolase activity	5	6,10
Translation factor activity	4	4,88
Transferase activity	3	3,66
Transporter activity	3	3,66
kinase activity	3	3,66
Nucleic acid binding	2	2,44
Motor activity	2	2,44
DNA binding	2	2,44
RNA binding	2	2,44
Sequence specific DNA binding transcription factor activity	1	1,22
Cytoeskeletal protein binding	1	1,22
Receptor binding	1	1,22
Lipid binding	1	1,22
Metal ion binding	1	1,22
Zinc ion binding	1	1,22
Total	82	100

From the annotated 82 molecular processes, most of them were grouped with a binding function (57.32%), by far the most repeated function, followed by structural molecule (9.76%) and catalytic activity (7.32%) (Table 4).

Table 4. Molecular processes groups that resulted when the molecular processes were grouped into larger units based on the similarities of the processes. (N: number of biological processes in the group; %: percentages).

Molecular processes groups	**N**	**%**
Binding	47	57,32
Structural molecule	8	9,76
Catalytic activity	6	7,32
Hydrolase activity	5	6,10
Translation	4	4,88
Transport	3	3,66
Others	9	10,98
Total	82	100,00

From the 56 sequences that had an alignment in the BLASTX, 40 were polymorphic in Atlantic Bluefin tuna and monomorphic in albacore. After introducing the sequences containing those 40 SNPs into the ORF Finder, the following results were observed: 20 synonymous SNPs, 4 non-synonymous SNPs and there rest of the SNPs (16) did not appear in an open reading frame (Table 5).

Table 5. Those SNPs that were and were not located within an open reading frame and among those located in an open reading frame Synonymous and non-synonymous SNPs. The non-synonymous include an amino acid change.

Contig name_ SNP locus	**ORF Finder**	**SNP type (aa substitution)**
92628_3851	No	
74975_1591	No	
74714_552	No	
74714_737	No	
74714_929	No	
74714_993	No	
74714_1068	Yes	Synonymous
68656_5309	Yes	Non-synonymous (Q -> H)
67565_7966	Yes	Synonymous
63575_3883	Yes	Synonymous

Contig name_ SNP locus	ORF Finder	SNP type (aa substitution)
56436_2296	Yes	Synonymous
56414_1700	Yes	Non-synonymous (K ->)
50048_1747	Yes	Synonymous
46200_8178	Yes	Synonymous
4613_3618		
43186_223	Yes	Synonymous
41319_565	Yes	Synonymous
36284_1658	Yes	Synonymous
34678_2934	No	
33532_1358	Yes	Synonymous
33239_1303	No	
28360_3462	Yes	Non-synonymous (P -> T)
26540_1032	No	
23848_5959	Yes	Synonymous
21470_885	Yes	Synonymous
21470_954	No	
20740_4954	No	
18871_2901	Yes	Synonymous
14837_11318	Yes	Synonymous
12328_390	Yes	Synonymous
121813_4643	No	
121382_2923	Yes	Synonymous
113966_10824	Yes	Synonymous
113966_7437	Yes	Synonymous
113966_5180	Yes	Synonymous
111946_1396	No	
110330_6560	Yes	Non-synonymous (E -> V)
108905_2168	No	
107346_1369	No	
104772_860	Yes	Synonymous

A summary of the results in each step is shown in Figure 4.

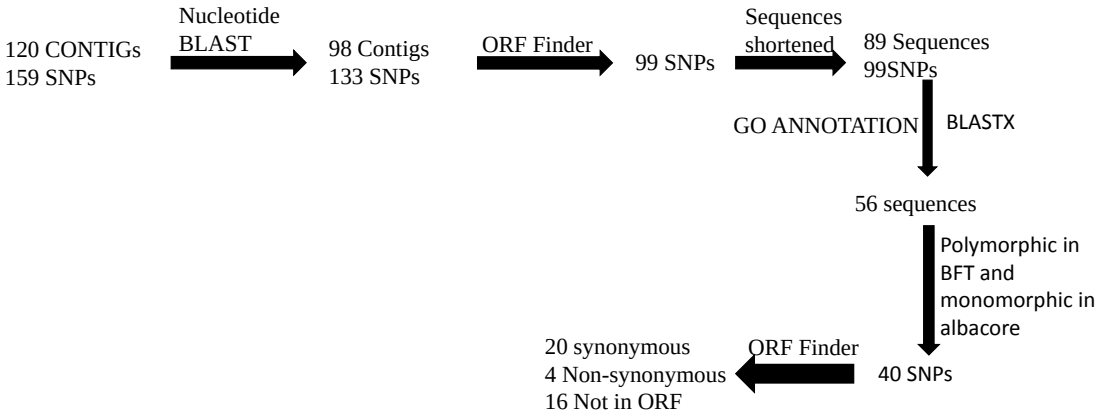

Figure 4. The resulting SNPs and contigs after each step of the analysis.

DISCUSSION

Fifty-six sequences had a significant alignment in the BLASTX and within them 40 SNPs were polymorphic in Bluefin tuna and monomorphic in albacore. Among these SNPs, 24 were located in an open reading frame and 20 were synonymous SNPs and four non-synonymous. This means that the variation seen in those four SNPs might be underlying an adaptation. Those 20 synonymous SNPs could be important as well because a change in the nucleotide can result in an alteration of a protein. In order, to know if these variations underlie any type of adaptation or specialization, further analysis should be done considering local adaptations for different populations and a characterization of each population would be necessary.

In this study, orthologous genes from the muscle transcriptome in *Thunnus thynnus* were analyzed using different bioinformatics tools. In multicellular organisms, cells have nearly the same genome and therefore, the same genes. However, not all genes are active in each cell, that is, gene expression changes in different cells. These variations underlie the differences shown by different cells and tissues. A transcriptome is the part that is transcribed into RNA molecules and it is a very small part of the genetic code, just about the 5%. It seems that the proportion of transcribed sequences that are non-protein-coding is greater in more complex organisms that is not all the transcriptome is translated into proteins (Adams, 2008). However, all the sequences of the transcriptome used in this study are protein coding.

From the initial 120 contigs 98 contigs (%81.67) were aligned whereas 18.33% of the contigs did not have a significant alignment (cutoff E value > 10^{-5}). In order to interpret this, first it is important to know that in the NCBI database all the sequences of the zebrafish (*Danio rerio*),

which is the model organism for fish (that is, the whole genome has been sequenced), are present, both annotated and unannotated genes. Additionally, as previously said, all the transcriptome sequences in this study are protein coding. Thus, even if 22 contigs had no significant alignment this does not mean that they are not part of a protein coding sequence but that the sequence they are located in, has not been annotated yet. Therefore, this is likely to be due to the fact that this is a sequence that just appears in *T. thynnus* and *T. alalunga* or it could be a sequence they share with some other fish whose genome have not been fully analyzed.

Within those 98 contigs there were 133 SNPs. From those just the SNPs that were located in an open reading frame were selected, 98 SNPs. Then, the sequences in which these 98 SNPS were located were shortened and introduced in the BLASTX. From the 89 sequences introduced in the Blast2GO, 56 had a significant alignment when the BLASTX was run. It is very likely that this is because the amino acid sequences have not been identified for those nucleotide sequences, that is, the genes for those sequences are unannotated at the protein level. Even in humans, the most analyzed organism, a certain percentage of the genes are unannotated (Pertea et al., 2010) so it is logical that some genes in *T. thynnus* are not annotated. There are two possibilities: one is that the proteins just appear in *T. thynnus* although this in unlikely because they were aligned in the previous step and the other and most probable is that they share this protein with other teleost fish although the protein remains unknown.

The GO terms that have been given to most of the 56 sequences that had a significant alignment (cutoff value > 10^{-6}) in the BLASTX. However, 15 of those 56 (26.78%) sequences were not given a biological process and 17 (30.36%) sequences were not given a molecular process (annexes table 4). From the initial 120 contigs, after the sequences were shortened, just 56 (46.67%) had an alignment with protein sequences. From this 56 sequences, 45 were annotated with a biological process or molecular process, that is 37.5% of the initial sequences and 34 sequences had annotations for both biological and molecular process that is 28.33% of the initial sequences. This annotation efficiencies are a little bit lower to those obtained in some other studies with marine fish, which are around 45% (Coppe et al., 2010; Mu et al., 2010; Palstra et al., 2013; Xiang et al., 2010), but this can be due to the absence of a sequenced *T. thynnus* genome.

Observing the results, the bioinformatics tools used in this study have proven to be successful as 81.67% of the sequences were annotated as genes. Blast2GGO has been used in a couple of other studies of fish populations where positive results were achieved (De Wit et al., 2010; Huth and Place, 2013).

It is very interesting to compare the percentages of each biological and molecular processes with those from other studies in order to compare the gene expression in *T. thynnus* and other

organisms, specially fish, as this could be very informative in terms of adaptation to a given environment or habitat. However, it is important to remember that the transcriptome varies from tissue to tissue so transcriptomes of different tissues should not be compared between different organisms. This is because if you try to analyze an adaptation or the differences of gene expression in different organisms based on transcriptomes for different tissues, this can lead false interpretations and conclusions. When two transcriptomes form different tissues within an individual are compared, this can be used to make a comparison of gene expression in different tissues as it has been done in a study where transcriptomes form the muscle, ovaries and testis of the giant freshwater prawn were analyzed (Jung et al., 2011). Another study has analyzed the transcriptome of the rainbow trout (Palstra et al., 2013).

It is interesting to compare the annotation of the transcriptome of different organisms and in the case of the Atlantic Bluefin tuna and albacore, it is of special interest to compare it to other fish. In this study, the most abundant biological processes were metabolic processes with a relative abundance of 28% that is very similar to the 25% observed in a study of the European hake (Milano et al., 2011). However, when biological processes are compared in both studies it can be seen that the most abundant biological processes in the study of the muscle transcriptome of the European hake were cellular processes, which were not observed in the present study. Another important difference is observed in developmental processes that in the present study showed a relative abundance of 24.46% while in the study of the hake it is just the 9%. In regulation processes, a difference is also observed, in the case of the hake they constituted the 6% and in the two *thunnus* species, the 9.35%. In the response to stimulus, a difference is seen too, from a 5.76% in the present study to a 2% of the biological processes in the hake. In the case of transport, these processes constituted the 5.76% of the biological processes of this two *thunnus* species but was not present in the case of the hake. These differences can be due to the fact that in the case of the two *thunnus* species only those genes that are conserved were analyzed while in the case of the hake a more general analysis of the transcriptome was made. It can also be inferred that cellular processes are not very conserved, as they are not present in the case of the conserved genes.

It is also interesting to compare the molecular processes with the same study of the European hake (Milano et al., 2011). In the case of the hake, the most abundant molecular process was binding the same as in the present study but the relative abundance was different, in the case of the two species in the genus *thunnus* it constituted a 57.32% and in the hake it was a 42%. A major difference is observed in the case of catalytic processes these two studies. In the present study, it constituted the 7.32% whereas in the hake it was the 34%. The structural molecule activity is quite similar in both studies, 9.76% in the present study and 11% in the European hake. As it happens in the case of biological processes, these differences are observed because the genes of the present

study are those that are conserved in the Atlantic Bluefin tuna and albacore and thus, it seems that the catalytic processes are not highly conserved. On the other hand, binding is highly conserved.

In order to have a deeper understanding of the importance of these biological and molecular processes, the pathways of these processes should be studies more deeply. This can be an interesting analysis for future studies. .

GO terms define gene sets. However, their importance in a biological process or pathway in a dataset is determine not just by how many genes are identified but also by their relative abundance ratio considering the total number of genes involved in each pathway. This is why precise mapping of metabolic pathways becomes necessary and the usefulness of pathway analysis software is largely increasing (Khatri et al., 2012). Up to now, there is only one fish metabolic network available, MetaFishNet, a public online tool (Li et al., 2010). As the metabolic pathways very important when studying biological processes, further studies can be directed toward analyzing these pathways for *T. thynnus* and *T. alalunga*. As this study has identified some genes and their annotation has been done, this can lead to an analysis to infer the pathways for those genes and processes in the future.

The gene annotation of conserved genes performed in this study, offers new tools for further studies of the Atlantic Bluefin tuna and albacore. This allows having a better understanding of the ecology and evolution of these species and this is important in order to establish a proper conservation plan and an appropriate management.

REFERENCES

Adams, J. 2008. Transcriptome: connecting the genome to gene function. *Nature Education* 1(1): 19.

Allendorf, F. W., Hohenlohe, P. A. and Luikart, G. 2010. Genomics and the future of conservation genetics. *Nature Reviews Genetics* 11(10): 697-709.

Allendorf, F. W., Luikart, G. H. and Aitken, S. N. 2013. *Conservation and the Genetics of Populations*. Wiley-Blackwell, West Sussex.

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3): 403-410.

Amato, G., Ryder, O. A., Rosenbaum, H. C. and DeSalle, R. 2008. *Conservation Genetics in the Age of Genomics*. Columbia University Press, New York.

Avise, J. C. 2010. Perspective: conservation genetics enters the genomic era. *Conservation Genetics* 11(2): 665–669.

Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* 282(5396): 2022-2028.

Collette, B. B. 2001. Scombridae. In: K. E. Carpenter and V. Niem (eds), *The Living Marine Resources of the Western Central Pacific*. FAO, Rome.

Collette, B., Amorim, A.F., Boustany, A., Carpenter, K.E., de Oliveira Leite Jr., N., Di Natale, A., Die, D., Fox, W., Fredou, F.L., Graves, J., Viera Hazin, F.H., Hinton, M., Juan Jorda, M., Kada, O., Minte Vera, C., Miyabe, N., Nelson, R., Oxenford, H., Pollard, D., Restrepo, V., Schratwieser, J., Teixeira Lessa, R.P., Pires Ferreira Travassos, P.E. and Uozumi, Y. 2011a. *Thunnus thynnus*. In: IUCN 2014. IUCN Red List of Threatened Species. Version 2014.1. <www.iucnredlist.org>. Downloaded on **23 June 2014**.

Collette, B., Acero, A., Amorim, A.F., Boustany, A., Canales Ramirez, C., Cardenas, G., Carpenter, K.E., Chang, S.-K., de Oliveira Leite Jr., N., Di Natale, A., Die, D., Fox, W., Fredou, F.L., Graves, J., Guzman-Mora, A., Viera Hazin, F.H., Hinton, M., Juan Jorda, M., Minte Vera, C., Miyabe, N., Montano Cruz, R., Masuti, E., Nelson, R., Oxenford, H., Restrepo, V., Salas, E., Schaefer, K., Schratwieser, J., Serra, R., Sun, C., Teixeira Lessa, R.P., Pires Ferreira Travassos, P.E., Uozumi, Y. and Yanez, E. 2011b. *Thunnus alalunga*. In: *IUCN 2014. IUCN Red List of Threatened Species.* Version 2014.1. <www.iucnredlist.org>. Downloaded on 26 March 2014.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. and Robles, M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18): 3674-3676.

Coppe, A., Pujolar, J. M., Maes, G. E., Larsen, P. F., Hansen, M. M., et al. 2010. Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics* 11(1): 635.

De Wit, M., Keil, D., van der Ven, K., Vandamme, S., Witters, E. and De Coen, W. 2010. An integrated transcriptomic and proteomic approach of characterizing estrogenic and metabolic effects of 17 α-ethinylestradiol in zebrafish (*Danio rerio). General and Comparative Endocrinology* 167(2): 190-201.

DeWoody, J. A., Bickham, J. W., Michler, C. H., Nichols, K. M., Rhodes, G. E., and Woeste, K. E. 2010. *Molecular approaches in natural resource conservation and management*. Cambridge University Press, New York.

Fassler, J. and Cooper, P. BLAST Glossary. 2011 Jul 14. In: *BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008.* Available from: http://www.ncbi.nlm.nih.gov/books/NBK62051/ Downloaded on 20 May 2014.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic Biology* 19(2): 99-113.

Frankham, R. 2010. Challenges and opportunities of genetic approaches to biological conservation. *Biological Conservation* 143(9): 1919–1927.

Fromentin, J. M. and Powers, J. E. 2005. Atlantic bluefin tuna: population dynamics, ecology, fisheries and management. *Fish and Fisheries* 6(4): 281-306.

Huth, T. J. and Place, P. 2013. *De novo* assembly and characterization of tissue specific transcriptomes in the emerald notothe, *Trematomus bernacchii. BMC Genomics* 14: 805.

Jung, H., Lyons, R. E., Dinh, H., Hurwood, D. A., McWilliam, S., et al. 2011. Transcriptomics of a Giant Freshwater Prawn (*Macrobrachium rosenbergii*): *De Novo* Assembly, Annotation and Marker Discovery. *PLoS ONE* 6(12): e27938.

Khatri, P., Sirota, M. and Butte, A. J. 2012. Ten years of pathway analysis: current approaches and outstanding challenges: P*LoS Computational Biology* 8(2)**:** e1002375.

Laconcha, U., Iriondo, M., Manzano, C., Markaide, P., Montes, I., Zarraonaindia, I., Velado, I., Bilbao, E., Goñi, N., Santiago, J., Pardo, M. A., Domingo, A., Delgado, A., Karakulak, S., Oray, I., Brophy, D., Arrizabalaga, H. and Estonba, A. (unpublished). New nuclear and mitochondrial SNP markers reveal albacore (*Thunnus alalunga*, Bonn.) population genetic structure between ocean basins and provide a basis for the sustainable management of the species.

Li, S., Pozhitkov, A., Ryan, R. A., Manning, C. S., Brown-Peterson, N. and Brouwer, M. 2010. Constructing a fish metabolic network model. *Genome Biology* 11(11)**:** R115.

MacKenzie, B. R., Mosegaard, H. and Rosenberg, A. A. 2009. Impending collapse of bluefin tuna in the northeast Atlantic and Mediterranean. *Conservation Letters* 2(1): 26-35.

Madden, T. The BLAST Sequence Analysis Tool. 2002. Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, (eds). *The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002*. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/ Downloaded on May 28 2014.

Maguire, J. J. 2006. *The state of world highly migratory, straddling and other high seas fishery resources and associated species*. FAO, Rome.

Majkowski, J. 2007. *Global fishery resources of tuna and tuna-like species*. FAO, Rome.
Milano, I., Babbucci, M., Panitz, F., Ogden, R., Nielsen, R. O., et al. 2011. Novel Tools for Conservation Genomics: Comparing Two High-Throughput Approaches for SNP Discovery in the Transcriptome of the European Hake. *PLoS ONE* 6(11): e28008.

Mu, Y., Ding, F., Cui, P., Ao, J., Hu, S., et al. 2010. Transcriptome and expression profiling analysis revealed changes of multiple signalling pathways involved in immunity in the large yellow croaker during Aeromonas hydrophila infection. *BMC Genomics* 11(1): 506.

Mushegian, A. R., Garey, J. R., Martin, J. and Liu, L. X. 1998. Large-scale taxonomic profiling of eukaryotic model organisms: A comparison of orthologous proteins encoded by humans, fly, nematode, and yeast genomes. *Genome Research.* 8(6): 590-598.

Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma R. K. and Hedrick P. W. 2010. Conservation genetics in transition to conservation genomics. *Trends in Genetics* 26(4): 177–187.

Palstra, A. P., Beltran, S., Burgerhout, E., Brittijn, S. A., Magnoni, L. J., et al. 2013. Deep RNA Sequencing of the Skeletal Muscle Transcriptome in Swimming Fish. *PLoS ONE* 8(1): e53171.

Pertea, M. and Salzberg, S. L. 2010 .Between a chicken and a grape: estimating the number of human genes. *Genome Biology* 11(5): 206.

Primmer, C. R. 2009. From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences* 1162(1): 357–368.

Rubin , G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., et al. 2000. Comparative genomics of the eukaryotes. *Science* 287(5461): 2204-2215.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. 1997. A genomic perspective on protein families. *Science* 278(5338): 631-637.

Tatusov, R. L., Galperin, M. Y, Natale, D. A., and Koonin, E. V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28(1): 33-36.

Wheelan, S. J., Boguski, M. S., Duret, L. and Makalowski, W. 1999. Human and nematode orthologs-lessons from the analysis of 1800 human genes and the proteome of *Caenorhabilis degans*. *Gene* 238(1): 163-170.

Wheeler, D. L., Barret, T., Benson, D. A., Bryant, S. H., Canese, K., Chetverin, V., Church, D. M. and DiCuccio, M. 2007. Database resources of the Nacional Center for Biotechnology Information. *Nucleic Acid Research* 35(1): D5-D12.

Worm, B. and Tittensor, D.P. 2011. Range contraction in large pelagic predators. *Proceedings of the National Academy of Sciences* 108(29): 11942-11947

Xiang, L. X., He, D., Dong, W-R., Zhang, Y-W. and Shao, J-Z. 2010. Deep sequencing-based transcriptome profiling analysis of bacteria-challenged *Lateolabrax japonicus* reveals insight into the immune-relevant genes in marine fish. *BMC Genomics* 11(1): 472.

ANNEXES

Table 1: The predicted genes afther the nucleotide BLAST was run and if the SNP is located within an open reading frame (RF).

Contig name_ SNP position	PREDICTED GENE	RF
1807_1464		No
4613_3618	chromosome sequence corresponding to linkage group 1	Yes
9675_583	myosin light chain 2 (MCL) mRNA	Yes
12328_390	guanidinoacetate N-methyltransferase-like, mRNA	Yes
14148_3956	ATP synthase subunit beta, mitochondrial-like, mRNA	Yes
14837_11318	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 10, mitochondrial-like, mRNA	Yes
14837_6079		Yes
16636_17148	26S proteasome complex subunit DSS1-like, mRNA	Yes
17065_727		No
17183_11293	fibroleukin-like, transcript variant (X2), mRNA	Yes
17183_11458		Yes
18452_6637		No
18871_2901	mRNA for thioredoxin-interacting protein, complete cds/ mRNA for pituitary adenylate cyclase activating polypeptide receptor 1A (pacap1A gene)	Yes
20740_4954	60S ribosomal protein L3 putative mRNA, pseudogene cds/ribosomal protein L3-like mRNA, complete cds	Yes
21470_885	heat shock protein beta-11-like, mRNA	Yes
21470_954		Yes
22149_2163	microtubule-associated protein futsch-like, mRNA	Yes
22149_3252		Yes
22149_3357		Yes
22848_5312	synemin-like, mRNA/ Desmuslin putative mRNA	Yes
23724_3855		No
23848_5959	60S ribosomal protein L13 mRNA	Yes
24268_592		No

Contig name_ SNP position	PREDICTED GENE	RF
24563_365	basigin-like transcript variant X2, mRNA	Yes
25320_4698		No
25408_3114	NADH-ubiquinone oxidoreductase 75 kDa subunit, mitochondrial-like, mRNA	Yes
25556_5115	Gamma-aminobutyric acid receptor-associated protein-like 1 putative mRNA	Yes
25556_5226		Yes
26181_3568	uncharacterized, transcript variant X2, misc_RNA/elastin-like, mRNA	Yes
26540_1032	sestrin-1-like, transcript variant X2, mRNA	Yes
26806_1288	extended synaptotagmin-2-A-like, transcript variant X4, mRNA	Yes
27156_337	G0/G1 switch protein 2-like, mRNA	Yes
27279_824	ribosomal protein S11-like mRNA/ 40S ribosomal protein S11 / mRNA for very low-density lipoprotein receptor precursor (vtgr gene)	Yes
27555_4907	chromosome sequence corresponding to linkage group 1, top part,/ Myosin regulatory light chain 2, smooth muscle isoform putative mRNA/ putative transient receptor protein 2 mRNA	Yes
27555_6076		Yes
27997_2787	TGF-beta-inducible nuclear protein 1 putative mRNA/ ribosome biogenesis protein NSA2 homolog (LOC102208086), mRNA	Yes
28360_3462	Zebrafish DNA sequence from clone CH1073-207P11 in linkage group 21, complete sequence	Yes
29338_778	60S ribosomal protein L21 mRNA	Yes
30754_3141	apolipoprotein O-like, mRNA	Yes
31952_2466	beta-defensin gene/ reproduction regulator 1 (rr1) mRNA	Yes
33239_1303	40S ribosomal protein S18 mRNA	Yes

Contig name_ SNP position	PREDICTED GENE	RF
33532_1358	gene for fast skeletal myosin heavy chain isoform mMYH-11	Yes
33532_713		Yes
34220_1259	cofilin-2-like, mRNA	Yes
34509_3915		No
34678_2934	calcium-bindin g mitochondrial carrier protein Aralar1-like, transcript variant X2, mRNA	Yes
35533_1235		No
35548_2427	60S ribosomal protein L32-like, mRNA	Yes
35902_6157	galectin-1 (Glec-1) mRNA/ Beta-galactoside-binding lectin putative mRNA,	Yes
36284_1658	Procollagen C-endopeptidase enhancer 2 precursor putative mRNA/ gamma-aminobutyric acid receptor beta subunit gene, partial cds; 55kd erythrocyte membrane protein (P55), synaptic vesicle-associated integral membrane protein (VAMP-1), procollagen C-proteinase enhancer protein (PCOLCE) genes, complete cds; glucose repression mediator protein (GRMP) gene, partial cds	Yes
36819_2169	UBX domain-containing protein 2A-like, transcript variant X2, mRNA	Yes
38002_1999	skeletal muscle atypical fast troponin T isoform 1 mRNA	Yes
38882_106	ribosomal protein S12 (rps12), transcript variant 2, mRNA/ 40S ribosomal protein S12-like, mRNA	Yes
40138_3002	very-long-chain (3R)-3-hydroxyacyl-[acyl-carrier protein] dehydratase 1-like, transcript variant X2, mRNA/	Yes
40492_2374	phosphoglycerate kinase 1-like, mRNA	Yes
40814_3109		No
41319_565	elongation factor 2-like, mRNA	Yes
42859_5670	gene fore parvalbumin beta/ parvalbumin beta-1 gene	Yes
43186_223	calponin-1-like, transcript variant X2, mRNA	Yes

Contig name_ SNP position	PREDICTED GENE	RF
43186_3117		Yes
45092_1603	betaine--homocysteine S-methyltransferase 1-like, mRNA	Yes
45275_16096	Signal recognition particle 14 kDa protein putative mRNA, complete cds/ neuroglobin-like (LOC100708898), mRNA	Yes
45879_1468	hydroxyacyl-Coenzyme A dehydrogenase mRNA, partial cds/ trifunctional enzyme subunit beta, mitochondrial-like, mRNA	Yes
45879_1926		Yes
46200_8178	clone VMRC26-161K14, complete sequence	Yes
47828_2787	fructose-1,6-bisphosphatase isozyme 2-like, mRNA	Yes
50048_1747	cytochrome b-c1 complex subunit 6, mitochondrial-like, mRNA	Yes
51302_3262		No
51363_1363	alcohol dehydrogenase class III mRNA, complete cds/ alcohol dehydrogenase class-3-like, mRNA	Yes
51571_1098	cellular retinoic acid-binding protein 2-like, mRNA	Yes
51571_840		Yes
52293_1028		No
52955_2028	phospholipid hydroperoxide glutathione peroxidase, mitochondrial-like, misc_RNA/ glutathione peroxidase 4b mRNA, complete cds	Yes
54349_347	synaptopodin 2-like protein-like, transcript variant X5, mRNA/ myozenin-1-like (LOC102293347), transcript variant X2, mRNA	Yes
54349_593		Yes
55307_234		No
56414_1700	aspartate aminotransferase, mitochondrial-like, mRNA	Yes
56414_5302		Yes
56436_2296	calsequestrin-1-like, mRNA	Yes
56470_1385		No
56538_3410	telethonin-like, mRNA	Yes

Contig name_ SNP position	PREDICTED GENE	RF
56694_1169		No
56694_2291		No
56694_2694		No
57535_921	collagen alpha-1(VI) chain-like, mRNA	Yes
59131_9207	calcium uptake protein 3, mitochondrial-like, transcript variant X5, mRNA	Yes
59655_1477	creatine kinase S-type, mitochondrial-like, mRNA/ creatine kinase mitochondrial isoform mRNA, complete cds; nuclear gene for mitochondrial product	Yes
59655_1718		Yes
60098_2850		No
60098_899		No
60564_6105	fast/white muscle troponin T embryonic isoform mRNA, complete cds// skeletal muscle fast troponin T embryonic/larval isoform mRNA, complete cds	Yes
60806_190	thrombospondin-2-like, transcript variant X4, mRNA	Yes
60806_388		Yes
60806_663		Yes
61631_1423	tropomyosin alpha-4 chain-like, transcript variant X8, mRNA	Yes
63177_1281	ribosomal protein L4 mRNA, partial cds/ 60S ribosomal protein L4-A-like, mRNA	Yes
63177_445		Yes
63575_3883	troponin C mRNA, complete cds	Yes
65377_1455	high mobility group protein B1-like, mRNA	Yes
65377_677		Yes
66298_2101	uncharacterized, mRNA	Yes
66558_2260		No
67565_7966	cytochrome b-c1 complex subunit Rieske, mitochondrial-like, mRNA	Yes
68656_5309	myeloid leukemia factor 1-like, transcript variant X3, mRNA	Yes
71405_1764	caldesmon-like, mRNA	Yes
74456_78	mRNA, clone: BRF 27-E3, induced by treatment of LPS	Yes

Contig name_ SNP position	PREDICTED GENE	RF
74714_1068	nebulin-like, mRNA	Yes
74714_552		Yes
74714_737		Yes
74714_929		Yes
74714_993		Yes
74975_1309	collagen alpha-1(VI) chain-like, mRNA	Yes
74975_1591		Yes
74975_607		Yes
75178_2809		No
76689_116	mRNA for ribosomal protein LP0, complete cds	Yes
77539_117	chromosome sequence corresponding to linkage group 1, bottom part, complete sequence	Yes
81111_1351		No
81830_85	partial mRNA for elongation factor 1-alpha (ef1a gene)	Yes
89148_263	lipoma-preferred partner homolog, transcript variant X3, mRNA	Yes
90889_248	myosin light chain 2 (MCL) mRNA, complete cds	Yes
90889_62		Yes
92628_3851	alpha actin (alpha-cardiac actin3) gene, complete cds	Yes
96746_1758		No
103799_348	, cation transport regulator-like 1, mRNA/from clone CH211-244P18 in linkage group 20 Contains the gene for a novel protein similar to vertebrate delta-like 4 (Drosophila) (DLL4), the vps18 gene for vacuolar protein sorting protein 18, the gene for a novel protein containing an SNF2 family N-terminal domain and a Helicase conserved C-terminal domain, the gene for a novel protein similar to vertebrate ras homolog gene family, member V (RHOV), the gene for a novel protein containing a ChaC-like protein domain and three CpG islands, complete sequence	Yes
103841_1366	ATP synthase, H+ transporting, mitochondrial F1 complex, delta subunit	Yes

Contig name_ SNP position	PREDICTED GENE	RF
103947_3022	NADP-dependent malic enzyme, mitochondrial-like, mRNA	Yes
104143_2111	ribosomal protein L8 mRNA/ 60S ribosomal protein L8 mRNA	Yes
104143_4113		Yes
104143_4176		Yes
104549_3250	eukaryotic translation initiation factor 5A-1-like, transcript variant X2, mRNA	Yes
104772_860	mRNA for slow myosin heavy chain (MyoHC-A4 gene)/gene for fast skeletal myosin heavy chain isoform mMYH-11	Yes
105438_909		No
106376_2098	calponin-1-like, transcript variant X2, mRNA	Yes
106376_2446		Yes
107346_1369	gene for fast skeletal myosin heavy chain isoform mMYH-	Yes
107794_2694	trinucleotide repeat-containing gene 18 protein-like , transcript variant X1, mRNA	Yes
107794_2938		Yes
108670_3117	microtubule-associated protein 4-like, mRNA	Yes
108905_2168	polycystin-1-like, transcript variant X3, mRNA	Yes
110330_6560	surf3 and surf6 genes and partial surf1 gene	Yes
111523_3798	delta-sarcoglycan-like, transcript variant X3, mRNA	Yes
111523_4842		Yes
111523_5475		Yes
111946_1396	M-protein, striated muscle-like, transcript variant X4, mRNA/myomesin-1-like (LOC102301909), mRNA	Yes
111946_5188		Yes
113966_10824	phosphoglucomutase-1-like, transcript variant X2, mRNA	Yes
113966_5180		Yes
113966_7437		Yes
114055_2490		No
114273_1588		No
114273_2811		No
114893_4017	protein S100-A11-like mRNA	Yes

Contig name_ SNP position	PREDICTED GENE	RF
114893_4160		Yes
115192_1122	xin actin-binding repeat-containing protein 2-like (LOC102204054), mRNA	Yes
116699_1486	transcription elongation factor B polypeptide 2-like, mRNA	Yes
121382_2923	eukaryotic translation initiation factor 3 subunit G (eTIF3) mRNA	Yes
121509_7273	PDZ and LIM domain protein 4-like, mRNA	Yes
121813_4643	Nicotinamide riboside kinase 2 putative mRNA/ muscle-specific beta 1 integrin binding protein 2 mRNA	Yes

Table 2: Predicted gene, the e value of the BLASTX hit, the reading frame (RF) for each SNP and if the SNP was aligned (SNP) after the BLAST2GO was run.

Contig name_ SNP locus	Predicted Gene	e value	RF	SNP
92628_3851	actin, alpha, cardiac muscle	0	2	No
81830_85	eukaryotic translation elongation factor 1 alpha	1,28E-25	2	Yes
74975_1591	collagen alpha-1(VI) chain-like	1,00E-78	-3	No
74714	neb protein	4,77E-28	-1	Yes
68656	myeloid leukemia factor 1-like isoform X3	8,63E-12	-2	Yes
67565	cytochrome b-c1 complex subunit Rieske, mitochondrial-like	8,15E-39	-3	Yes
65377_1455	high mobility group protein B1-like	1,15E-16	3	Yes
65377_677	high mobility group protein B1-like	1,22E-26	1	Yes
63575	troponin C, skeletal muscle-like	1,01E-24	1	Yes
59655_1718	creatine kinase S-type, mitochondrial	3,98E-14	1	Yes
59655_1477	creatine kinase S-type, mitochondrial-like	5,69E-08	3	Yes
56538_3410	telethonin-like	3,95E-20	-2	Yes

Contig name_ SNP locus	Predicted Gene	e value	RF	SNP
56436_2296	calsequestrin 1	1,65E-42	-2	Yes
56414_1700	aspartate aminotransferase, mitochondrial precursor	3,47E-16	-2	Yes
54349_347	miozenin-1-like	1,47E-30	3	Yes
50048_1747	cytochrome b-c1 complex subunit 6, mitochondrial-like	4,36E-27	3	Yes
46200_8178	pyrubate kinase PKM-like	7,60E-90	1	Yes
4613_3618	glycerol-3-phosphate dehydrogenase	3,84E-15	-2	Yes
43186_3117	calponin-1-like isoform X2	2,67E-18	-3	Yes
43186_223	calponin-1-like	1,80E-22	-1	Yes
41319_565	elongation factor 2-like	6,64E-74	-3	Yes
38882_106	40S ribosomal protein S12	1,06E-17	2	Yes
36284_1658	procollagen C-endopeptidase enhacer 2-like	6,56E-20	3	Yes
34678_2934	calcium-binding mitochondrial carrier protein Aralar1-like isoform X2	2,94E-35	-3	Yes
33532_1358	myosin heavy chain, fast skeletal muscle-like	1,08E-45	-3	Yes
33239_1303	40S ribosomal protein S18-like	3,63E-17	3	Yes
30754_3141	apolipoprotein O-like	8,77E-72	2	No
29338_778	60S ribosomal protein L21	1,10E-18	1	Yes
28360_3462	T-complex protein 1 subunit zeta-like isoform 1	1,50E-19	3	No
27555_6076	myosin regulatory light polypeptide 9-like	5,02E-29	3	No
27555_4907	myosin regulatory light polypeptide 9-like isoform X1	5,51E-40	3	Yes
26540_1032	unnamed protein	2,11E-12	2	Yes
23848_5959	60S ribosomal protein L13_like	1,74E-25	-3	Yes
22848_5312	synemin-like	6,51E-148	-3	Yes
22149	microtubule-associated protein futsch-like	2,40E-154	-1	Yes (2)
21470	heat shock protein beta-11-like	4,90E+00	-1	Yes (2)
20740_4954	60S ribosomal protein l3	4,65E-22	2	Yes
18871_2901	thioredoxin-interacting protein	1,23E-77	-2	Yes

Contig name_ SNP locus	Predicted Gene	e value	RF	SNP
14837_11318	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex subunit 10, mitochondrial-like	5,57E-16	3	Yes
14148_3956	AF419161 1 F1 ATP synthase bete subunit	2,03E-83	-1	No
12328_390	guanidinoacetate N-methyltransferase	1,06E-30	-2	Yes
121813_4643	unnamed protein product	5,22E-16	-2	Yes
121382_2923	eukatyotic translation initiation factor 3 subunit G-like	4,54E-22	3	Yes
115192_1122	xin actin-binding repeat-containing protein 2-like isoform X2	0,00E+00	2	Yes
113966_10824	phosphoglucomutase -1-like	2,17E-42	-3	Yes
113966_7437	phosphoglucomutase-1-like isoform X2	4,78E-37	-2	Yes
113966_5180	unnamed protein product	2,15E-21	-3	Yes
111946_1396	M-protein, striated muscle-like isoform X5	1,47E-25	-2	Yes
110330_6560	ribosomal protein L7, partial	1,13E-23	-2	Yes
108905_2168	polycystin-1-like isoform X3	5,50E-41	3	No
107346_1369	myosin heavy chain, fast skeletal muscle-like	4,51E-72	-2	Yes
104772_860	myosin heavy chain, fast skeletal muscle-like	9,92E-67	2	Yes
104549_3250	eukaryotic translation initiation factor 5A-1-like isoform 1	1,38E-33	-3	Yes
104143	60S ribosomal protein L8	2,02E-33	3	Yes
103841_1366	ATP synthase subunit delta, mitochondrial-like isoform 1	1,94E-15	-1	Yes
103799_348	cation transport regulator-like protein 1-like	1,47E-67	-1	Yes

Table 3. Biological process and molecular process for the predicted gene for the sequence where these SNPs are located after BLAST2GO was run.

Contig name_ SNP locus	Biological process	Molecular process
92628_3851	anatomical structure morphogenesis; multicellular organismal development; regulation of biological process; cell death; cell differentiation; cytoeskeleton organization; transport; nucleobase-containing compound metabolic process; catabolic process	cytoeskeletal protein binding; hydrolase activity; nucleotide binding
81830_85	Translation; nucleobase-containig compound metabolic process; catabolic process.	Translation factor activity; nucleic acid binding; hydrolase activity; nucleotide binding
74975_1591	Multicellular organismal development	
74714	Multicellular organismal development; transport; cell differentiation; regulation of biological process; cytoeskeleton organization	Structural molecule activity; actin binding
68656		
67565	ion transport; generation of precursor metabolites and energy	binding; catalytic activity; transporter activity
65377_1455	cellular component organization; regulation of biological process; multicellular organismla development; DNA metabolic process; response to external stimulus; response to stress; cell death; cell differentiation; response to abiotic stimulus; reproduction	protein binding; sequence-specific DNA binding trasncription factor activity; DNA binding
65377_677	Response to stress; response to abiotic stimulus; multicellular organismal development;	DNA binding
63575		actin binding; calcium ion binding
59655_1718	Metabolic process	Kinase activity; nucleotide binding
59655_1477	Metabolic process	Kinase activity; nucleotide binding
56538_3410	Cellular component organization; cell differentiation	
56436_2296		Calcium ion binding
56414_1700	Metabolic process; catabolic process; biosynthetic process; transport	Transferase activity; lipid binding; protein binding; binding;

Contig name_ SNP locus	Biological process	Molecular process
54349_347		
50048_1747	transport; generation of precursor metabolites and energy	catalytic activity; transporter activity
46200_8178	metabolic process; generation of precursor meabolites and energy; carbohydrate metabolic process; catabolic process	Kinase activity; binding
4613_3618	Carbohydrate metabolic process; catabolic process	nucleotide binding; catalytic activity; protein binding
43186_3117	cytoeskeleton organization	actin binding
43186_223	regulation of biological process; cytoeskeleton organization	actin binding; protein binding
41319_565	Translation; nucleobase-containig compound metabolic process; Catabolic process; embryo development	translation factor activity; nucleic acid binding; protein binding; hydrolase activity; nucleotide binding
38882_106	translation	structural molecule activity
36284_1658		
34678_2934	single-organism transport	
33532_1358		actin binding; nucleotide binding; motor activity
33239_1303	nucleobase-containingcompound metabolic process; catabolic process; cellular component organization; translation; multicellular organismal development; regulation of biological process; cell cycle; reproduction; viral process	structural molecule activity; RNA binding
30754_3141		
29338_778	translation	structural molecule activity; hydrolase activity
28360_3462	protein metabolic process; reproduction; cell recognitionnucleotide binding; protein binding	
27555_6076	Anatomical structure morphogenesis; cellular component organization; cell differentiation; regulation of biological process	Structural molecule activity; calcium ion binding
27555_4907	Anatomical structure morphogenesis; cellular component organization; regulation of biological process	Calcium ion binding
26540_1032		
23848_5959	regulation of biological process; cell cycle; translation	structural molecule activity

Contig name_ SNP locus	Biological process	Molecular process
22848_5312	structural molecule activity	cytoskeleton organization
22149		
21470	response to stress	
20740_4954	translation; multicellular organismal development	structural molecule activity
18871_2901		
14837_11318	nucleobase-containing compound metabolic process	transferase activity; nucleotide binding
14148_3956	ion transport; nucleobase-containing compound metabolic process; biosinthetic process; catabolic process	
12328_390	reproduction; regulation of biological process; growth; anatomical structure morphogenesis; multicellular organismal development; biosynthetic process; metabolic process	transferase activity
121813_4643	multicelluar organismal development; cellular component organization; signal transduction	
121382_2923	cellular component organization; translation; regulation of biological process	translation factor activity; protein biding; nucleotide binding
115192_1122	actin cytoeskeleton organization	metal ion binding; actin binding; zinc ion binding
113966_10824	carbohydrate metabolic process; catabolic process; generation of precursor metabolites and energy; biosynthetic process	calcium ion binding; binding; catalytic activity
113966_7437	generation of precursor metabolites and energy; carbohydrate metabolic process; catabolic process; biosynthetic process	binding; catalytic activity
113966_5180	response to abiotic stimulus; generation of precursor metabolites and energy; arbohydrate metabolic process; catabolic process; cellular homeostasis; biosynthetic process	binding; catalytic activity
111946_1396		
110330_6560		
108905_2168		

Contig name_ SNP locus	Biological process	Molecular process
107346_1369		actin binding; nucleotide binding; motor activity
104772_860		nucleotide binding
104549_3250	translation; regulation of biological process; cellular proten modification process; cellular component organization	translation factor activity; binding
104143	translation; reproduction; viral process; cellular component organization; protein transport; nucleobase-containing compound metabolic process; catabolic process	RNA binding; structural molecule activity
103841_1366	generation of precursor metabolites and energy; nucleobase containing compound metabolic process; ion transport; biosynthetic process;	hydrolase activity; transporter activity
103799_348	cell death; cell differentiation; multicellular organismal development; regulation of biological process; protein metabolic process; signal transduction; response to stress; response to biotic stimulus	receptor binding