Arab Universities Journal of Agricultural Sciences
(2023) 31 (1) 119-130
Website: http://ajs.journals.ekb.eg

Partial Gap-Filling of the Nile Tilapia (*Oreochromis niloticus*) Draft Genome

Hala M Zoghly, Mohamed A Rashed, Mahmoud Magdy*
Genetic Dept, Fac of Agric, Ain Shams Univ, P.O. Box 68, Hadayek Shoubra 11241, Cairo, Egypt

*Corresponding author: m.elmosallamy@agr.asu.edu.eg

https://doi.org/10.21608/AJS.2023.172400.1501

Received 2 November 2022; Accepted 3 December 2022

Keywords:

Freshwater fish,
Gene annotation,
Genomic gaps,
Sanger sequencing,
Gene prediction

Abstract: Nile tilapia is a freshwater fish of African origin, with productive and economic characteristics affecting global fish markets. The *Oreochromis niloticus* reference genome has a diploid set of 22 linkage groups (LGs) in addition to ungrouped sequences. A total of 42,622 genes have been identified, but 550 genomic gaps may include more. Our study focused on some of these genomic gaps, where appropriate primers were designed and then multiplied through the polymerase chain reaction (PCR) technique. From local samples, DNA was extracted and amplified with the new primers. Successful amplicons were sequenced and annotated using available bioinformatics tools. Five assessed sequences were annotated, of which three were newly predicted in *O. niloticus*, as mucin 1-like, and KLR genes, and SATB region. The other two sequences belonged to grid2 and trpm7 which were previously predicted. Although individual attempts to identify genomic gaps are not feasible in filling the large lack of information on the Nile tilapia genome, a good number and valuable new information has been reached. However, the following plan showed many technical problems, more time, effort and cost that could be avoided by suggesting the application of another technique, such as whole-genome sequencing, mapping, and assembly.

1 Introduction

Nile tilapia, or *O. niloticus* was firstly named by Linnaeus in 1758, and it belongs to the genus Oreochromis of the Cichlidae family. It is endemic to the Nile River Basin and lakes in Africa and is found especially around the tropics and subtropics (Trewavas 1983), but it has been introduced into more than 50 countries on all continents except Antarctica. It represents a good vertebrate model in evolutionary mechanisms studies due to its close relationship to haplochromine cichlids, which have undergone rapid speciation in East Africa. Tilapias are a good fish model in physiological studies due to their ability to adapt to environmental stresses (Hulata 2001). It is worth noting that Nile tilapia is known globally as a fish of difficult conditions, e.g., poor quality water, and although it prefers warm water, it can hold out slightly cold water (Rahman et al 2021). It can also adapt to a wide range of diets, low oxygen levels, and salinity (El-Sayed 2020). It is one of the first domesticated fish to be cultured in Egypt about 5000 years ago, from which it headed to the rest of the world for production or farming (Vajargah 2021). Nile tilapia is characterized by many desirable characteristics, such as a pleasant taste and ease of breeding and production at low cost, all of which led to making it one of the most widespread tilapia fish after carp worldwide (Prabu et al 2019). Nile tilapia is one of the fish that can be used as a highly nutritious fish meal that can replace other animal meat meal sources (Jim et al 2017). All the good qualities of this species have influenced its use in fish farming. It is adapted to most production strategies such as intensive, semi-intensive, and extensive systems, making it the third most important farmed fish species in the world with a production rate that increases annually to reach 4.5 million tons in 2018 (FAO 2022). Despite all the good qualities of tilapia, the attention is still not very strong due to its problems such as speed of reproduction, early maturity, crowding, and intense and rapid competition for food, and consequently the low quality of the water in which it is raised quickly, ending in poor quality of fish. However, efforts are made to solve these problems such as the methods used to produce monosex fish (Wokeh and Orose 2021).

In support of the efforts made to solve the problems facing this fish, it is necessary to provide its full genomic information as an integral part of the reasons for its behavior in the surrounding environment. Therefore, the state of its genome is a matter of knowledge, as ideas have been repeated in setting cognitive limits and building blocks to draw the genome picture of this fish species (Conte et al 2017). Although filling genomic gaps of Nile tilapia has not been a major goal of studies based on this species since ancient times, it is an implicit goal that is being achieved, thus the beginnings were mostly based on linkage genetic maps as in Kocher et al (1998) study in which 41 haploid embryos were used to draw a genetic that built on 162 DNA markers linked in a final map that spans 704 Kosambi cM consisting of 30 linkage groups representing 22 chromosomes. With the advent of second-generation sequencing technology Lee et al (2005) were able to construct a second-generation linkage map of the F2 progeny resulting from the cross between Nile tilapia and blue tilapia, including a study of 21 gene-based markers and 525 microsatellites. The map extended to 1311 cM, producing 22 linkage groups. With the development of second-generation sequencing platforms and the possibility of high accuracy of whole genome sequencing, attempts have been made to assemble the Nile tilapia genome. Among the genome assemblies available at the National Center for Biotechnology Information (NCBI) are Orenil1.1, Nile Tilapia GIFT, XX_nile, and O_niloticus_UMD_NMBU (NCBI 2022). The latest Nile tilapia genome assembly "O_niloticus_UMD_NMBU" is accepted as a reference genome so it was relied upon in the current study, and its major characteristics are 22 linkage groups in addition to mitochondrial DNA and other ungrouped sequences, which have a total length of 1,005,664,923 bp, 550 spanned gaps, 2,460 scaffolds (N50: 38,839,487 bp) and 40.5 % GC content (NCBI 2022). The incomplete information of this reference genome has made it a rich research material, especially with the rise of computational biology approaches, where there are now many servers, databases, algorithms, programs, and even web-based tools. Gene prediction programs enrich the identification of functional genes without the need to conduct laboratory experiments, which have time and cost that are difficult to confront in different cases (Goel et al 2013). There are two major types of gene prediction tools; the first is based on similarity research as a basic local alignment search tool (BLAST), and the second is based on signals and gene structure-based search, called Ab initio gene prediction programs e.g. FGENESHM GENSCAN, and AUGUSTUS (Wang et al 2004). The amplification of a genetic region by PCR and then determining its sequence by the Sanger method represents a successful and effective strategy in completing what is missing from the genetic information of the organism since ancient times (Innis et al 1988, Dorit et al 2001). So, the thought of applying the same idea to fill the gaps in Nile tilapia is not far away. Although the previous strategy is expensive and requires time and effort, it is still effective in its desired goal (Wilting et al 2022). Since the field is still available and interesting to know the secrets of the Nile tilapia genome, in particular the Egyptian version, the objective of the current study became inevitable, to fill and annotate several nuclear genomic gaps as soon as. The general steps were: i) designing specific primers for randomly selected genomic gaps, ii) conducting polymerase chain reactions to amplify these gaps; and finally, identifying their sequences by the Sanger sequencing method and annotating newly obtained sequences as an attempt to fill these gaps. Interestingly, the aim of genomic gap filling is the beginning of what will be completed in future studies.

2 Material and Methods

2.1 Sample Collection and DNA Extraction

Protocol (I) of Li et al (2015) with some minor modifications according to the conditions of the study was applied to extract DNA from scales of fresh dead fish obtained from the Al-Sharkawiia channel, Egypt. Scales were collected, washed, and dried the weighted as 0.05 g and put in clean Eppendorf per replicate. Five hundred µl of lysis I (100 mM Tris-HCL, pH 8.0; 50 mM EDTA, pH 8.0; 1 mM $CaCl_2$; 0.5 % (w/w)

SDS) were added to each replicate. Samples were incubated at boiling point in a water bath for 30 min then they were left to cool down at room temperature. Two hundred µl of lysis II (0.2 M EDTA, pH 8.0, 0.6 M NaCl, 6 % (w/w) SDS) and 10 µl proteinase K (20 mg/ml) were added to each tube and then incubated at 65°C for 30 min. Two hundred µl ammonium acetate (7.5 M) were added to each tube and then incubated at 4°C for 20 min. All samples were centrifuged at 12,000 g / 15 min. The supernatant was transferred into the new 1.5 Eppendorf, and then an equal volume of pre-cooled absolute ethanol was added. The tubes were slightly shaken and then incubated at -20°C for one hour. Samples were centrifuged at 12,000g / 15min, and then the supernatant was discarded. The pellets were cleaned with 70% ethanol and then dried on air. After complete disposal of the ethanol, 50 µl of sterile water was added to the samples to dissolve DNA. The obtained DNA was tested on 1 % Agarose gel.

2.2 Primer Design and Polymerase Chain Reaction (PCR) Conduction

Fifty-four genomic gaps (with an average of two or three gabs per linkage group) were randomly selected from O_niloticus_UMD_NMBU (GCF_001858045.2) reference genome. Thus, there is an attempt in the current study to fill the genomic gaps at a rate of 54/550. Fifty-four pairs of primers were designed by Geneious Prime 2022.1 software (https://www.geneious.com), but the current study focused on only five genomic gaps **(Table 1)**. Primer pairs were synthesized by Invitrogen, UK. Primer3web version 4.1.0 web-based tool (https://primer3.ut.ee/) was used for testing hairpin formation and self-annealing of the designed primers. Standard PCR was performed in a total volume of 20 µl of PCR mixture that consisted of 10 µl of amaR OnePCR™ (GeneDireX, Inc., United State), 0.5 µl reverse primer, 0.5 µl forward primer, 1 µl of DNA template and nuclease-free water up to 20 µl. Stratagene MX3000 P (Agilent Technologies) machine was programmed as follows: initial denaturation at 94°C / 5 min, repeated 35 cycles of 94°C / 30 sec for the denaturation phase, 50°C-68°C / 30 sec for the annealing phase, and 72°C / 2 min for extension phase. The final extension phase was at 72°C/10 min. The obtained PCR products were tested on 1.5% Agarose gel.

2.3 DNA Sequencing and DATA Analysis

PCR products were purified and labeled by Big Dye terminator V.3.1 Cycle Sequencing kit (Applied Biosystems, Inc.) and introduced to bidirectional sequencing using ABI 3730 automated Sanger sequencer (Macrogene, Inc.). BioEdit 7.2 software (Hall 1999) was used to qualify, clean, and align forward and reverse sequences obtained for each sample. Assessed sequences were annotated depending on some bioinformatic tools and databases. To increase the accuracy of the analysis of the refined sequences, a genomic sequence before and after desired genomic gaps was recovered from the reference genome (GCF_001858045.2). This genomic region was about 20,000 bp in the standard case in addition to the sequence itself, so the genomic region contains the sequence under study and approximately 10,000 bases before and after its chromosomal position.

Putative DNA/ protein motifs were retrieved through MEME SUIT 5.4.1 (Bailey et al 2015), but sequence annotation mainly depended on *ab initio*-based gene prediction software and homology search-based tools. Gene prediction programs such as GENSCAN (Burge and Karlin 1997), FGENEH v2.6 (Salamov and Solovyev 2000), and AUGUSTUS v3.3.3 (Stanke et al 2008) were used with the default parameters, except that the model organism was *O. niloticus* in case of FGENESH, and it was Zebrafish (*Danio rerio*) in case of AUGUSTUS. Putative promoter sequences were predicted by using Promoter v2.0 servers (Knudsen 1999) to check whether there are functional elements adjacent to the filled gaps or not. However, homology-based search tools such as BLAST with the cutoff e-value = 1 x 10 $^{-15}$ for BlastN and 1 x 10^{-6} for BlastP and BlastX were used (Altschul et al 1990). BlastN search was applied against the nucleotide collection (nt/nr) database with the determination of Oreochromini (taxid:1315725) as a preference organism according to obtained results. BlastX and BlastP were applied against non-redundant protein sequences (nr) and UniProtKB/Swiss-Prot (SwissProt) also with the determination of Oreochromini (taxid:1315725) as the preference organism.

Sequencing and gene prediction led to the assumption that there are 8 outputs for each sequence; the sequence itself (S1), the genomic region surrounding the sequence as much as possible, including the filled gap (S2), predicted CDS by GENSCAN (S3), FGENESH (S4), and AUGUSTUS (S5), and predicted peptides by GENSCAN (P1), FGENESH (P2), and AUGUSTUS (P3). DNA sequences (S1, S2, S3, S4, and S5) were searched by BlastN and BlastX, while predicted peptides (P1, P2, and P3) were submitted

Table 1. Gaps under study and their related primers, and linkage groups

GapID	Primer name	5'-seq-3'	Linkage group	Position
Gap 1	**Gap1F**	AGCAGCACAGAAGCAGGATAGC	LG1	836,755-836,854
	Gap1R	CACTGCACACACAGATATCCCC		
Gap 2	**Gap2F**	ATGGTTAGAGCGCCACCTAGTG	LG1	28,534,576-28,534,675
	Gap2R	TGGAAAATTGGCACAGAGCTTC		
Gap 3	**Gap3F**	CCGCCACACACATCAAACACG	LG5	3,798,065-3,798,164
	Gap4R	CCTGGTCTGCAGGCTCACC		
Gap 4	**Gap5F**	GACATTGTCTGAAAGCCAGCA	LG10	57,91,280-57,91,379
	Gap5R	AGGTGAAAGCATGAGGACGTA		
Gap 5	**Gap6F**	AGGACTCAAAGGTCACCCCC	LG12	21,607,702-21,607,801
	Gap6R	GCGTACCCAGCCTTTTCTCC		

to Interpro (Blum et al 2021) to functionally analyze predicted protein sequences and SWISS-MODEL server (Waterhouse et al 2018) to recognize the available supposed structural protein models. For each sequence, all possible reading frames were retrieved by the NCBI ORF finder online tool (https://www.ncbi.nlm.nih.gov/orffinder/).

3 Results and Discussion

DNA with acceptable quality **(Fig 1a)** of several tissue samples of *O. niloticus* fish was used to amplify the five gaps using conventional PCR which yielded five specific bands that were purified and sequenced **(Fig 1b)**. Five genomic gaps could be partly filled; thus, the physical structure of only two genomic gaps is recognized while the rest were not, but the current study can be characterized by 5/54 success gab filling rate. It is possible to attribute the lack of quality of the sequences to technical problems or the presence of more than one copy of the desired sequence amplified with the same primer, so it is possible to have several partially different copies of the same sequence which limits the full annotation.

The *ab initio* gene prediction proved that most of the studied sequences were introns, this output was supported by the potential 3 to 15 open reading frames that were predicted, but most of them had no significant hits in protein databases. This result supports that the sequences may be introns or intergenic, while the idea of being ncRNA or repetitive sequence is not far. All sequences were predicted as parts of potential genes by GENSCAN and FGENESH gene prediction programs, which have already proven efficient in gene prediction (Li et al 2017), but AUGUSTUS could not detect that for most cases, although it was effective in its performance with other types of fish e.g., *Amphilophus citrinellus* and *Archocentrus centrarchus* (Xiong et al 2021) in addition to the Nile tilapia (Jiang et al 2019).

As for gap 1, the amplified length could not be determined due to unqualified sequencing output. It was predicted as an intron of a potential gene by GENSCAN and FGENESH, while AUGUSTUS resulted in its being part of the intergenic region. Search by BlastN against the nt/nr database, specifically Oreochromini (taxid:1315725) resulted in shared significant hits for all queries (S1, S2, S3, and S4). The significant hits varied between immune genes (KLR and MHC class), gene expression regulation genes (VASA gene for ATP-dependent RNA helicase DDX4, ATP-dependent DNA helicase PIF1-like (LOC109195818), and VASA gene), and detoxification gene (tbtbp). BlastX search of the S1 and S4 was not significant, while S2 and S3 showed significant similarity to nascent polypeptide-associated complex subunit alpha,

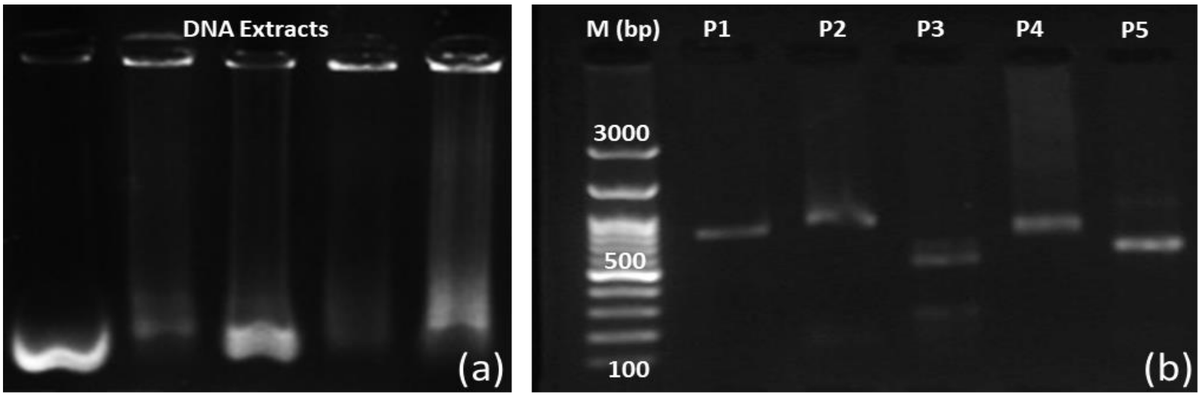

Fig 1. Extracted DNA from scales of freshly sampled Nile tilapia fish (a), and the PCR amplification of five final products (b)

muscle-specific form-like, and mucin-1-like. BlastP search of P2 resulted in no significant hits, while P1 showed the same result as the BlastX search. Interpro search resulted in no clear relation to a specific protein family, thus S1 showed integration into the not named family (PTHR32301), while P2 showed some signal peptides and non-cytoplasmic domain only. However, P1 resulted in the presence of a group of protein disorders. Protein structure modeling by the Swiss-Model server resulted in the homology of the P1 structure to Mariner Mos1 transposase (4r79.1.G). However, P2 appeared structural feature of CYTOCHROME C (1e86.1. A). Three potential open reading frames were detected by the ORF finder web-based tool for S1, but all of which have no significant hits. There were several potential promoter sequences predicted by the Promoter 2.0 server for S2, from which the promoter that was apart from predicted CDSs was about 1460 bp with a marginal expected score (0.634). Potential protein domains were retrieved by MEME-SUIT to ensure the relationship between predicted peptides and proposed proteins for mucin 1-like domains as shown in **Fig (2a)**, as one of the two closest hits (nascent polypeptide-associated complex subunit alpha, muscle-specific form-like, and mucin-1-like) for S1 and S2.

The first sequence was included in the potential gene predicted by GENSCAN, and this result was assured by the rest of the prediction tools as BLAST, Interpro, and Swiss-Model server that gave semi-similar significant hits. Unfortunately, the results were inconsistent and were not matched except in the case of searching for shared domains by MEME-SUIT which showed the presence of a protein shared between the expected peptides, mucin 1-like, and NASA gene. This result is the closest to accuracy because the protein is more accurate than DNA in identifying sequences. It is also in line with the results of BlastP for P1. Although P2 gave non-significant with BlastP search, it showed common domains with the two previously mentioned genes. The inconsistency of results and the lack of clear results may be due to the presence of problems in the sequence and the inability to know it completely, which means that even the results of the MEME-SUIT may change in the future with the knowledge of the complete sequence. If the previous interpretation is right, gap 1 can be part of the new gene copy as the available recorded copies of the mucin 1-like gene and nascent polypeptide-associated complex subunit alpha, muscle-specific form-like, as none were located on LG1 at all (NCBI 2022). All, this filled gap represents new addition in new copy gene prediction, which enriches the genetic background of Egyptian Nile tilapia. The newly discovered gene copies have great important action in fish life. Nascent polypeptide-associated complex subunit alpha, muscle-specific form-like has a role in the adaptation of living organisms to toxic chemicals, thus it has been proven that this gene is highly expressed in coral reef fish exposed to high concentrations of carbon dioxide (Tsang et al 2020). The mucin and mucin-like genes were discovered in humans and several organisms, not only fish, as these genes produce the mucin protein that enters the formation of mucus, which can protect the organism from biotic and abiotic challenges. Mucin and mucin-like genes also support the role of the digestive system in the digestion and absorption of food, and therefore their role is not limited to preventing the entry of pathogens into the cell only, but also promotes the growth of the fish to reach high sizes quickly (Aanyu et al 2018). This means that these genes can act as a strong indicator of the Egyptian environmental influences on the life of the fish in terms of its immune response or its growth, and we can also induce these genes in future studies to increase their expression to enhance the growth of the fish. Although GENSCAN and FGENESH produced that this sequence is an intron, it cannot be confirmed due to the incomplete filling of this genomic gap, the reason for the gap length which prevented the full amplification and sequencing of the gap, lead to less informative data.

As for gap 2, the amplified length could not be totally determined, but three used programs predicted it as an intron in a potential gene. BlastN, BlastX, and BlastP searches resulted in significant similarity to transient receptor potential cation channel subfamily M member 7 (TRPM7). Interpro search of all predicted peptides resulted in their integration in TRPM7 (IPR029601) and GO analysis resulted in their integration into three distinct categories. All predicted peptides showed significant similarity to TRMP7 (5zx5.1. A) when modeled by the Swiss-Model server. A potential promoter sequence with a highly likely prediction score (1.174) was detected at 300 bp apart from predicted CDSs. Eleven putative ORFs have resulted from ORF finder, all of which have no significant similarity of any protein when searched against a non-redundant protein database. However, only ORF 10 appeared significant similarity to the Ribulose bisphosphate carboxylase large chain. Probable protein domains that are shared between predicted peptides and TRPM7 were retrieved by the MEME-SUIT tool as shown in **Fig (2b)**.

The analysis of the second sequence was one of the lucky few cases as all three used gene prediction programs agreed that it was an intron in a potential gene. The results of using servers and database search also were in favor of unique output. All results of used tools shared the same output “TRPM7” in which the sequence is supposed to be included. TRPM7 is a gene that produces a membrane protein that forms a channel through which cations pass selectively, as this membrane channel participates in many important vital processes of the cell, especially during the development of embryos and the body's resistance to diseases, the role of the TRPM gene has been proven not only in humans but in vertebrates in general (Yee 2018). Another candidate evidence about being of gap 2 as part of TRPM7 gene is that this gap is indeed located at LG1 (accession: NC031965; position: 28534576-28534676) within the alleged gene that is located on LG1 (accession: NC031965; position: 28439676-28540542) (NCBI 2022). It's controversial the lack of significance of the ORFs of this sequence, but even the tenth one that had a significant hit was completely different from a previously desired gene, but this can be overlooked, as the sequence under study was predicted as an intron.

As for gap 3, the amplified length could not be determined. It was predicted as an intron of the potential gene by GENSCAN, while FGENESH showed its integration into poly A (a) and transcription binding site region (b) of adjacent potential genes. However, AUGUSTUS output that this sequence represented an intergenic region. BlastN search showed no significant hits for S1 and S4b that belonged to the second potential gene predicted by FGENESH, although S2, S3, and S4a showed significant similarity to Killer cell lectin-like receptors (KLR) and RAS-like family 12. BlastX search also resulted in no significant hits for the S1 and S4b, but the rest queries showed significant similarity to a group of uncharacterized proteins, serine/arginine repetitive matrix protein 1-like, and proteoglycan 4-like. All previous hits were the outputs of BlastP search of all predicted peptides, except the P2a that had no significant hits. Interpro search resulted in the integration of the sequence into a not named subfamily (PTHR14689:SF2) and included a non-cytoplasmic domain and several signal peptides. P1 showed integration into the SGNH hydrolase superfamily (SSF52266), indicating disorder prediction. P2a showed disorder prediction only, while P2b showed no significant prediction. Predicted peptides showed different protein models, thus P1 showed similarity to several different models, from which Carboxylesterase (7c23.1.A). This protein model belonged to the SGNH-hydrolase family esterase which supported the result of the Interpro search. P2a showed similarity to Adenylate kinase (2osb.2. A), while P2b showed similarity to Hemoglobin Alpha chain (1o1j.1.A). Only one potential promoter sequence was detected around 500bp apart from S4a with a high score (1.209). There were fifteen ORFs retrieved for the sequence, but all of which had no significant hits. Potentially shared protein domains were retrieved for predicted peptides and the most similar hits; proteoglycan4, RAS-like family 12, KLRs, and SGNH hydrolase superfamily (data not showed). The results were in favor of KLR genes due to the presence of two big, shared domains between predicted peptides and C-type lectin domain containing protein groups as shown in **Fig (3a)**.

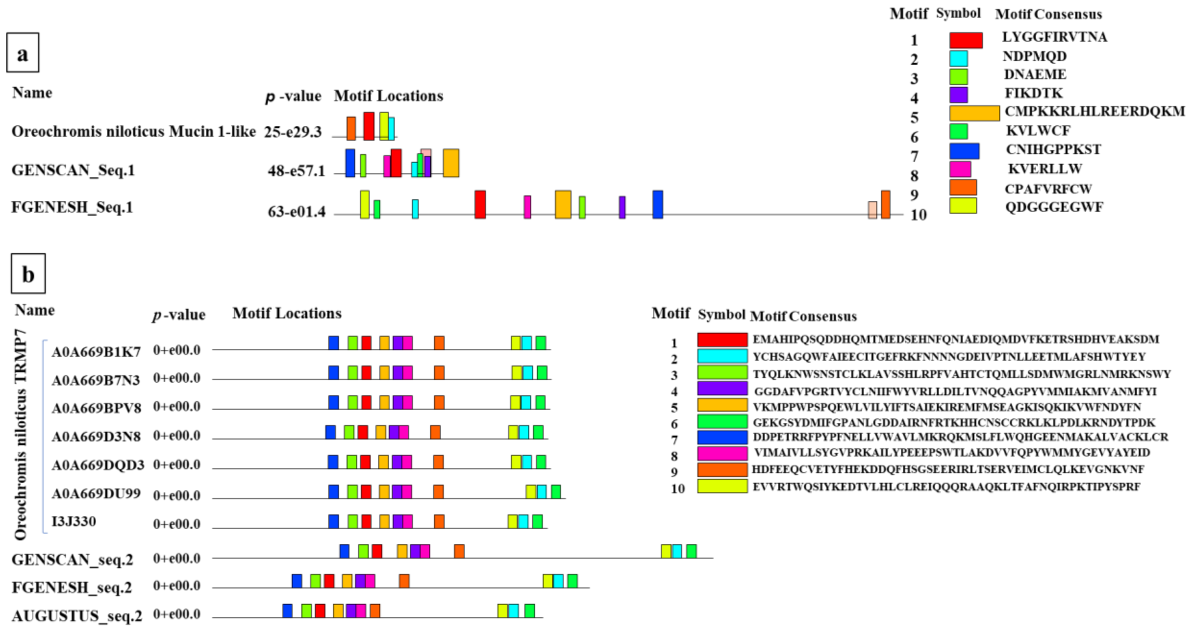
a
Name
p -value
Motif Locations
Oreochromis niloticus Mucin 1-like
25-e29.3
GENSCAN_Seq.1
48-e57.1
FGENESH_Seq.1
63-e01.4
Motif
Symbol
Motif Consensus
1 LYGGFIRVTNA
2 NDPMQD
3 DNAEME
4 FIKDTK
5 CMPKKRLHLREERDQKM
6 KVLWCF
7 CNIHGPPKST
8 KVERLLW
9 CPAFVRFCW
10 QDGGGEGWF
b
Name
p -value
Motif Locations
Oreochromis niloticus TRMP7
A0A669B1K7 0+e00.0
A0A669B7N3 0+e00.0
A0A669BPV8 0+e00.0
A0A669D3N8 0+e00.0
A0A669DQD3 0+e00.0
A0A669DU99 0+e00.0
I3J330 0+e00.0
GENSCAN_seq.2 0+e00.0
FGENESH_seq.2 0+e00.0
AUGUSTUS_seq.2 0+e00.0
Motif
Symbol
Motif Consensus
1 EMAHIPQSQDDHQMTMEDSEHNFQNIAEDIQMDVFKETRSHDHVEAKSDM
2 YCHSAGQWFAIEECITGEFRKFNNNNGDEIVPTNLLEETMLAFSHWTYEY
3 TYQLKNWSNSTCLKLAVSSHLRPFVAHTCTQMLLSDMWMGRLNMRKNSWY
4 GGDAFVPGRTVYCLNIIFWYVRLLDILTVNQQAGPYVMMIAKMVANMFYI
5 VKMPPWPSPQEWLVILYIFTSAIEKIREMFMSEAGKISQKIKVWFNDYFN
6 GEKGSYDMIFGPANLGDDAIRNFRTKHHCNSCCRKLKLPDLKRNDYTPDK
7 DDPETRRFPYPFNELLVWAVLMKRQKMSLFLWQHGEENMAKALVACKLCR
8 VIMAIVLLSYGVPRKAILYPEEEPSWTLAKDVVFQPYWMMYGEVYAYEID
9 HDFEEQCVETYFHEKDDQFHSGSEERIRLTSERVEIMCLQLKEVGNKVNF
10 EVVRTWQSIYKEDTVLHLCLREIQQQRAAQKLTFAFNQIRPKTIPYSPRF

Fig 2. The potential shared protein domains between predicted peptides of seq.1 and mucin-1-like (a), and potentially shared protein domains between predicted peptides of seq.2 and TRPM7 (b)

Since the third sequence has not been fully known, the results of gene predictors have varied between intron, polyadenine, transcription factor binding site, or intergenic region. Blast, Interpro, and SWISS-MODEL search results also conflicted. These conflicted outputs may be due to the low quality of the sequence itself, which negatively affected on sequence annotation process.

The low quality of amplification and sequencing may be due to technical problems or the gab itself is too long to be amplified by conventional PCR. Although the search based on the protein was the definitive means of identifying the sequence, in the present case it was not conclusive, but the presence of two common, clear, and large-sized protein domains between the proteins of this sequence and the domains of KLR genes made this target the closest. The presence of the non-cytoplasmic domain in the sequence itself supported the previous vision, which still needs strong foundations. If previous findings will be proven, this copy of the KLR gene will be a great discovery as this gene is a member of the gene family that has a great role to enhance the immune response, especially in overcoming the cancers of living organisms (Lee et al 2022). As for the role of KLRs in Nile tilapia, they have been proven to be upregulated after bacterial infection in tissues associated with the adaptive immune response, such as gills and spleen, as well as its role in activating T lymphocytes (Zhang et al 2021). The specific function of KLRs leads to their integration into bacterial resistance pathways. Therefore, these genes can play an important role in the adaptation of Nile tilapia to different biotic stresses, which means that we can benefit from upregulating these genes in any environment rich in pathogenic microbes.

As for gap 4, the amplified length was about 778 bp. It was predicted as an intron in potential gene by GENSCAN and FGENESH, while AUGUSTUS showed no gene structure was found for this sequence. BlastN search showed significant hits for all queries. S1 and S3 had unique but not shared hits, thus, S1 was similar to *O. niloticus* gastrula zinc finger protein XlCGF26.1 (LOC102080278), while S3 was similar to *O. niloticus* SPG21. S2 and S4 had multiple and shared hits that were *O. niloticus* zinc finger BED domain-containing protein 1-like (LOC112847726), and VASA gene for ATP-dependent RNA helicase DDX4, *O. niloticus* KLR, *O. niloticus* SPG21 and *O. niloticus* repetitive sequence (SATB). BlastX and BlastP showed no significant hits, and the same result was obtained from Interpro search for all queries; thus, no predicted domains or protein families except some signal peptides in the case of S1. Protein structure models that were predicted by the SWISS-MODEL server were different for both predicted peptides. Thus, P2 was close to the structural property of the related matrix-associated actin-dependent regulator of chromatin subfamily B member 1 (6ltj.1.M), while P1 shared the structure of UBP3-associated protein BRE5 (2qiy.1.A). Two potential promoter sequences were predicted with marginal scores (0.518-0.762) before initially predicted exons.

There were nine retrieved ORFs, but all of them had no significant hits. Since predicted protein or resulted from nucleotide translation, both were non-significant and the closest hit for the sequence under study was the repetitive sequence "SATB"; thus shared DNA motifs were searched between the previous hit and seq.4, as shown in **Fig (3b)**.

The fourth sequence didn't seem to have a good chance to be annotated as part of the gene, thus although its nucleotides were fully known, their genetic role could not be determined. BlastN search was only the valuable tool that resulted in different patterns of hits, which were highly significant, but with very little query coverage, so the ability to trust these hits is low. The agreement of the rest of the research tools on nothingness refuted that this sequence may be the intergenic region of the unexpressed DNA region. Also, the probability of the presence of the promoter was less, which reduced the possibility of the studied sequence being included in a potential gene but did not make it impossible. Search for potential DNA motifs between the sequence itself and the repetitive sequence (SATB) led to their convergence, reinforcing the previous results. SATB is a long monomer (up to 1.9 kb) non-coding repetitive DNA sequence present in up to hundreds of thousands of copies in the Nile tilapia genome (Tao et al 2021). Fish satellite DNA sequences have been proven to have a great role in sex chromosome differentiation, and chromosome rearrangements, they also could be useful in comparative and chromosome evolution studies, especially in Nile tilapia (Ferreira and Martins 2008). If the fourth sequence is already part of SATB, this means that the current gap might be included in the centromeric region of the related chromosome, which provides an opportunity for cytological studies with relative knowledge of the chromosomal structure by determining the position of the centromere of this linkage group, which contains the gap under study.

Instead of the previous interpretation, the sequence might be ncRNA, as the agreement of the sequencing analysis tools that it is not coding for a protein supports this idea as well. This investigation just needs more analysis and experimental evidence, not only prediction tools.

As for gap 5, the amplified length was 607 bp. Its analysis was feasible; thus, it was found to represent part of the *O. niloticus* grid2 “glutamate receptor, ionotropic, delta 2” (Gene ID: 100696181) by BlastN and BlastX search using a wide genomic region surrounding the sequence itself. Unfortunately, it is part of one of its introns located in 58,800-59,407 of the detected gene. Although there was no point in using the rest of the tools, for assurance gene predictors and ORFs were applied. FGENESH produced the sequence as an intron in a potential gene, while GENSCAN made the sequence an exon, but AUGUSTUS output it as an intergenic region. The prediction by gene predictors was far from grid2, so more explanation around these predictors is required. There were 8 potential ORFs, but also all of which had no significant hits. Shared protein domains for predicted peptides and glutamate receptors delta subfamily were shown in **Fig (3c)**.

The analysis of the fifth sequence was the luckiest and most reliable; thus BLAST search proved including this sequence in an already predictable gene, *O. niloticus* grid2 “glutamate receptor, ionotropic, delta 2”, despite being an intron with a length of 607 bp. Interestingly, desired sequence locus was already located inside grid2 paralog; thus its related genomic gap is in LG12 (position: 21607702-2160802), and the predicted grid2 is in LG12 (position: 21548808-22032498). Surprisingly, gene predictors were unable to prove this conclusion, except for FGENESH which proved that the sequence represents an intron, which is the only proven fact about the sequence. This makes each of the methods of predicting genes in comparison, which brings us closer to the truth, do we trust the methods based on similarity-based research or signals and gene structure-based search. According to the results of the current study, it is indispensable to use more than one strategy to predict the gene to be certain about the results obtained, although at the end, conducting laboratory experiments is still the most accurate. However, the result of the BLAST has already been confirmed with the location of studded gab, which assured that this gab is part of a Grid2 gene that already predicted. According to KEGG database **(**Kanehisa 2019**)**, Grid2 gene is integrated in neuroactive ligand-receptor interaction and long-term depression pathways, thus Grid2 gene encodes protein

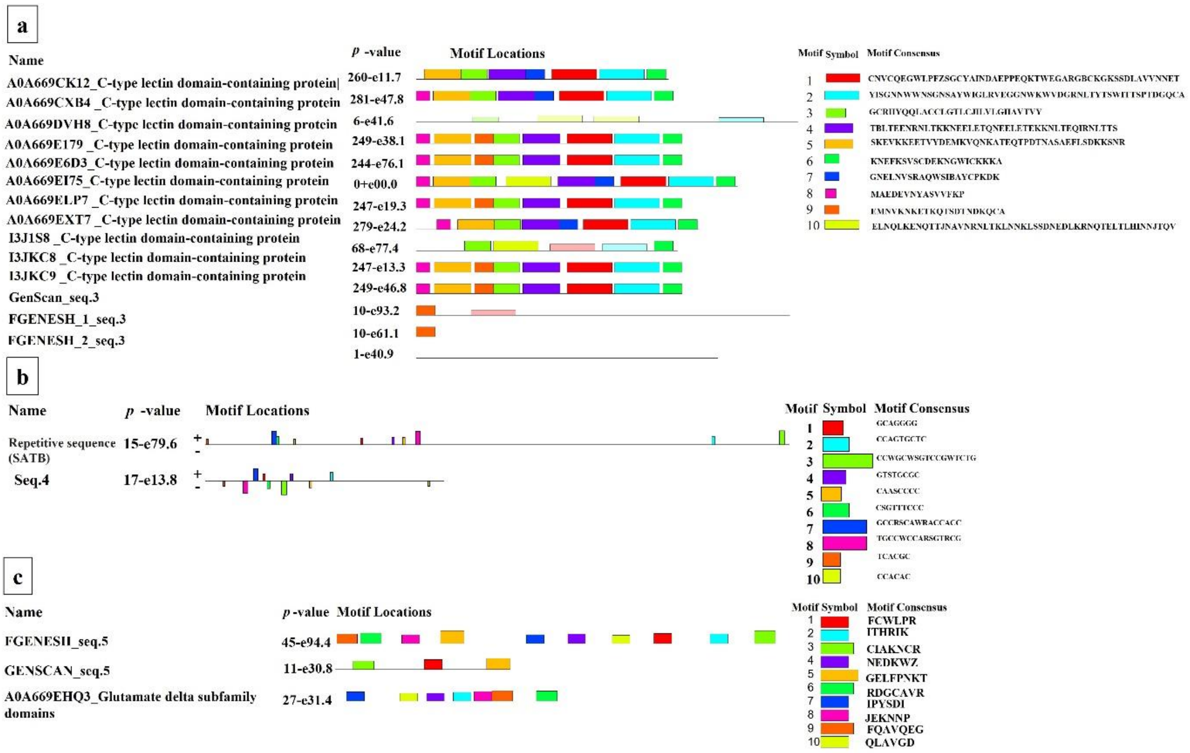
a
Name
p -value
Motif Locations
A0A669CK12_C-type lectin domain-containing protein
260-e11.7
A0A669CXB4 _C-type lectin domain-containing protein
281-e47.8
A0A669DVH8_C-type lectin domain-containing protein
6-e41.6
A0A669E179 _C-type lectin domain-containing protein
249-e38.1
A0A669E6D3_C-type lectin domain-containing protein
244-e76.1
A0A669EI75_C-type lectin domain-containing protein
0+e00.0
A0A669ELP7 _C-type lectin domain-containing protein
247-e19.3
A0A669EXT7 _C-type lectin domain-containing protein
279-e24.2
I3J1S8 _C-type lectin domain-containing protein
68-e77.4
I3JKC8 _C-type lectin domain-containing protein
247-e13.3
I3JKC9 _C-type lectin domain-containing protein
249-e46.8
GenScan_seq.3
10-e93.2
FGENESH_1_seq.3
10-e61.1
FGENESH_2_seq.3
1-e40.9
Motif Symbol Motif Consensus
1 CNVCQEGWLPFZSGCYAINDAEPPEQKTWEGARGBCKGKSSDLAVVNNET
2 YISGNNWWNSGNSAYWIGLRVEGGNWKWVDGRNLTYTSWITTSPTDGQCA
3 GCRHYQQLACCLGTLCJHLVLGHAVTVY
4 TBLTEENRNLTKKNEELETQNEELETEKKNLTEQIRNLTTS
5 SKEVKKEETVYDEMKVQNKATEQTPDTNASAEFLSDKKSNR
6 KNEFKSVSCDEKNGWICKKKA
7 GNELNVSRAQWSIBAYCPKDK
8 MAEDEVNYASVVFKP
9 EMNVKNKETKQTSDTNDKQCA
10 ELNQLKENQTTJNAVNRNLTKLNNKLSSDNEDLKRNQTELTLHINNJTQV
b
Name
p -value
Motif Locations
Repetitive sequence (SATB)
15-e79.6
+
-
Seq.4
17-e13.8
+
-
Motif Symbol Motif Consensus
1 GCAGGGG
2 CCAGTGCTC
3 CCWGCWSGTCCGWTCTG
4 GTSTGCGC
5 CAASCCCC
6 CSGTTTCCC
7 GCCRSCAWRACCACC
8 TGCCWCCARSGTRCG
9 TCACGC
10 CCACAC
c
Name
p-value Motif Locations
FGENESH_seq.5
45-e94.4
GENSCAN_seq.5
11-e30.8
A0A669EHQ3_Glutamate delta subfamily domains
27-e31.4
Motif Symbol Motif Consensus
1 FCWLPR
2 ITHRIK
3 CIAKNCR
4 NEDKWZ
5 GELFPNKT
6 RDGCAVR
7 IPYSDI
8 JEKNNP
9 FQAVQEG
10 QLAVGD

Fig 3. The potential shared protein domains between predicted peptides for gap3 and C-type lectin domains related to KLRs (a), the potential shared DNA motifs between gap4 and repetitive DNA sequence "SATB" (b), and the potential shared protein domains between predicted peptides for gap5 and glutamate delta subfamily domains (c)

that acts as a transmembrane receptor in brain cells, specifically in cerebellar Purkinje cells. This receptor is specific to glutamate amino acid that acts as a neurotransmitter in the neurotransmission process in the central nervous system. This receptor also plays a great role in brain development in the brain of vertebrates. Grid2 and Grid1 are members of the delta glutamate receptor family the ionotropic glutamate receptors family, which represents the fourth class of ionotropic glutamate receptors in addition to the rest three classes: N-methyl-D-aspartate (NMDA), α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA), and kainate (KA) receptors (Egbenya et al 2021). Therefore, this gene is a strong indicator of the quality of the environment surrounding the fish, especially during its early growth stages, which makes this gene important in future studies to reveal the extent of adaptation of fish in different environmental conditions.

4 Conclusion

It was not easy to reach the results of this study, it was hard work. Since the emergence of technical problems during DNA extraction and PCR calibration, as well as sequencing, there is still a need to further explore of the causes of these obstacles to find suitable solutions. Five genomic gaps could be sequenced after a long time of PCR standardization attempts. All studied genomic gaps are partially filled and they are probably part of genes; some were potentially new copies and some were already predicted genes that were recorded on NCBI. So, our findings promote Nile tilapia genome studies, which will benefit the productive quality of this fish. As for the tools used, each of them proved highly efficient in use, but computer-based predictions still lack accuracy. As for the gene prediction programs, both the GENSCAN and FGENESH competed for first place, as each of them proved efficient in their performance, which was confirmed by the similarity of the results, while AUGUS-

TUS failed to help in this matter. This may be due to the incompleteness of the Nile tilapia genome and its inclusion in the database used to search for this program. As for the carrying capacity, both FGENESH and AUGUSTUS were equal in their high ability to analyze large sizes up to multiples of gigabytes, while Genscan has a limited capacity, thus its maximum acceptable sequence length is 1 Mb. Finally, we recommend re-examining these genomic gaps by researching from a different point of view, while trying to avoid potential technical problems to verify the current study's findings. Further genetic studies are also recommended to investigate the molecular features of discovered genes. The related pathways and gene ontology also must be recognized, in addition to the relationship between genes and environmental adaptation of Nile tilapia in our Egyptian environment.

References

Aanyu M, Betancor Mónica B, Monroig O (2018) Effects of dietary limonene and thymol on the growth and nutritional physiology of Nile tilapia (*Oreochromis niloticus*). *Aquaculture* 488, 217-226. https://doi.org/10.1016/j.aquaculture.2018.01.036

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215, 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bailey TL, Johnson J, Grant CE, et al (2015) The MEME suite. *Nucleic Acids Research* 43, W39-W49. https://doi.org/10.1093/nar/gkv416

Blum M, Chang HY, Chuguransky S, et al (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* 49, D344-D354. https://doi.org/10.1093/nar/gkaa977

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268, 78-94. https://doi.org/10.1006/jmbi.1997.0951

Conte MA, Gammerdinger WJ, Bartie KL, et al (2017) A high-quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics* 18, 314. https://doi.org/10.1186/s12864-017-3723-5

Dorit RL, Ohara O, Hwang CBC, et al (2001) Direct DNA sequencing of PCR products. *Current Protocols in Molecular Biology* 56, 15.2.1-15.2.13. https://doi.org/10.1002/0471142727.mb1502s56

Egbenya DL, Aidoo E, Kyei G (2021) Glutamate receptors in brain development. *Child's Nervous System* 37, 2753-2758. https://doi.org/10.1007/s00381-021-05266-w

El-Sayed AFM (2020) Tilapia Culture. 2nd Edition, Academic Press, Alexandria, 348 p. https://doi.org/10.1016/C2017-0-04085-5

FAO (2022) The state of world fisheries and aquaculture- Towards Blue Transformation. Available online: https://www.fao.org/3/cc0461en/cc0461en.pdf (accessed on 30 October 2022).

Ferreira IA, Martins C (2008) Physical chromosome mapping of repetitive DNA sequences in Nile tilapia *Oreochromis niloticus*: Evidences for a differential distribution of repetitive elements in the sex chromosomes. *Micron* 39, 411-418. https://doi.org/10.1016/j.micron.2007.02.010

Goel N, Singh S, Aseri TC (2013) A review of soft computing techniques for gene prediction. *International Scholarly Research Notices (ISRN) Genomics* 2013, 191206. http://dx.doi.org/10.1155/2013/191206

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41, 95-98.

Hulata G (2001) Genetic manipulations in aquaculture: A review of stock improvement by classical and modern technologies. *Genetica* 111, 155-173. https://doi.org/10.1023/A:1013776931796

Innis MA, Myambo KB, Gelfand DH, et al (1988) DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proceedings of the National Academy of Sciences* 85, 9436-9440. https://doi.org/10.1073/pnas.85.24.9436

Jiang DL, Gu XH, Li BJ, et al (2019) Identifying a long QTL cluster across chrLG18 associated with salt tolerance in tilapia using GWAS and QTL-seq. *Marine Biotechnology* 21, 250-261. https://doi.org/10.1007/s10126-019-09877-y

Jim F, Garamumhango P, Musara C (2017) Comparative analysis of nutritional value of *Oreochromis niloticus* (Linnaeus), Nile tilapia, meat from three different ecosystems. *Journal of Food Quality* 2017, 714347. https://doi.org/10.1155/2017/6714347

Kanehisa M (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Science* 28, 1947-1951. https://doi.org/10.1002/pro.3715

Knudsen S (1999) Promoter 2.0: for the recognition of PolII promoter sequences. *Bioinformatics* 15, 356-361. https://doi.org/10.1093/bioinformatics/15.5.356

Kocher TD, Lee WJ, Sobolewska H, et al (1998) A genetic linkage map of a cichlid fish, the tilapia (*O. niloticus*). *Genetics* 148, 1225-1232. https://doi.org/10.1093/genetics/148.3.1225

Lee BY, Lee WJ, Streelman JT, et al (2005) A second-generation genetic linkage map of tilapia (Oreochromis spp.). *Genetics* 170, 237-244. https://doi.org/10.1534/genetics.104.035022

Lee LJ, Hassan N, Idris SZ, et al (2022) Differential Regulation of NK Cell Receptors in Acute Lymphoblastic Leukemia. *Journal of Immunology Research* 2022, 7972039 https://doi.org/10.1155/2022/7972039

Li XY, Liu XL, Ding M, et al (2017) A novel male-specific SET domain-containing gene setdm identified from extra microchromosomes of gibel carp males. *Science Bulletin* 62, 528-536. https://doi.org/10.1016/j.scib.2017.04.002

Li Y, Gul Y, Cui L, et al (2015) Comparative analysis of different protocols for extraction of DNA from fish scales of *Cyprinus carpio*. *Indian Journal of Biotechnology* 14, 382-387. https://nopr.niscpr.res.in/handle/123456789/33415

National Center for Biotechnology Information (NCBI) (2022) Assembly Information by Organism. Available at: https://www.ncbi.nlm.nih.gov/assembly/organism/8128/all.

Prabu E, Rajagopalsamy CBT, Ahilan B, et al (2019) Tilapia–An excellent candidate species for world aquaculture. A review. *Annual Research and Review in Biology* 31, 1-14. https://doi.org/10.9734/arrb/2019/v31i330052

Rahman ML, Shahjahan M, Ahmed N (2021) Tilapia farming in Bangladesh: Adaptation to climate change. *Sustainability* 13, 7657. https://doi.org/10.3390/su13147657

Salamov AA, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research* 10, 516-522. https://doi.org/10.1101/gr.10.4.516

Stanke M, Diekhans M, Baertsch R, et al (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24, 637-644. https://doi.org/10.1093/bioinformatics/btn013

Tao W, Xu L, Zhao L, et al (2021) High-quality chromosome-level genomes of two tilapia species reveal their evolution of repeat sequences and sex chromosomes. *Molecular Ecology Resources* 21, 543-560. https://doi.org/10.1111/1755-0998.13273

Trewavas E (1983) Tilapiine fishes of the genera Sarotherodon, Oreochromis and Danakilia. London, British Museum (Natural History), UK, 583 p. https://doi.org/10.5962/bhl.title.123198

Tsang HH, Welch Megan J, Munday PL, et al (2020) Proteomic responses to ocean acidification in the brain of juvenile coral reef fish. *Frontiers in Marine Science* 7, 605. https://doi.org/10.3389/fmars.2020.00605

Vajargah MF (2021) A review of the physiology and biology of Nile tilapia (*Oreochromis niloticus*). *Journal of Aquaculture and Marine Biology* 10, 244-246. https://medcraveonline.com/JAMB/JAMB-10-00328.pdf

Wang Z, Chen Y, Li Y (2004) A brief review of computational gene prediction methods. *Genomics, Proteomics and Bioinformatics* 2, 216-221. https://doi.org/10.1016/S1672-0229(04)02028-5

Waterhouse A, Bertoni M, Bienert S, et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* 46, W296-W303. https://doi.org/10.1093/nar/gky427

Wilting A, Nguyen TV, Axtner J, et al (2022) Creating genetic reference datasets: Indirect sampling of target species using terrestrial leeches as sample “collectors”. *Environmental DNA* 4, 311-325. https://doi.org/10.1002/edn3.256

Wokeh OK, Orose E (2021) Use of Dietary Phytochemicals as a control for Excessive Breeding in Nile Tilapia (*Oreochromis niloticus*): A review. *GSC Biological and Pharmaceutical Sciences* 17, 152-159. https://doi.org/10.30574/gscbps.2021.17.2.0336

Xiong P, Hulsey CD, Fruciano C, et al (2021) The comparative genomic landscape of adaptive radiation in crater lake cichlid fishes. *Molecular Ecology* 30, 955-972. https://doi.org/10.1111/mec.15774

Yee NS (2018) Transient Receptor Potential Cation Channel Subfamily M Member 7. In: Choi S (Eds), Encyclopedia of Signaling Molecules. Springer, Cham, pp 5649-5661. https://doi.org/10.1007/978-3-319-67199-4_101913

Zhang Y, Li K, Li C, et al (2021) An atypical KLRG1 in Nile tilapia involves in adaptive immunity as a potential marker for activated T lymphocytes. *Fish and Shellfish Immunology* 113, 51-60. https://doi.org/10.1016/j.fsi.2021.03.016