Journal of *Marine Science and Engineering*

MDPI

Article

Whole Genome Sequence of the Newly Prescribed Subspecies *Oreochromis spilurus saudii*: A Valuable Genetic Resource for Aquaculture in Saudi Arabia

Mohammed Othman Aljahdali [1,*], Mohammad Habibur Rahman Molla [1] and Wessam Mansour Filfilan [2]

[1] Biological Sciences Department, Faculty of Science, King Abdulaziz University, P.O. Box 80203, Jeddah 21589, Saudi Arabia; mrahmanmolla@stu.kau.edu.sa

[2] Department of Biological Sciences, Aljamom University College, Umm Al-Qura University, P.O. Box 2203, Makkah 25376, Saudi Arabia; wmfilfilan@uqu.edu.sa

* Correspondence: moaljahdali@kau.edu.sa

Abstract: Tilapia (*Oreochromis* spp.) have significant potential for aquaculture production around the world. There is an increasing demand among tilapia producers for strains with higher yields and for fish that can survive in highly saline water. Novel strains and consistent seedstock are critically important objectives for sustainable aquaculture, but for these required targets there is still not enough progress. Therefore, this study describes the genome sequence of *Oreochromis spilurus* to support the seawater culture of tilapia. The draft genome is 0.768 Gb (gigabases), with a scaffold N_{50} (the genome (50%) is in fragments of this length) of 0.22 Mb (megabases). The GC content is 40.4%, the heterozygosity rate is 0.35%, and the repeat content is 47.97%. The predicted protein-coding peptide encoded 51,642 and predicted 10,641 protein-coding genes in the *O. spilurus* genome. The predicted antimicrobial peptides were 262, bringing new hope for further research. This whole genome sequence provides new insights for biomedical and molecular research and will also improve the breeding of tilapia for high yields, resistance to disease, and adaptation to salt water.

Keywords: tilapia (*O. spilurus*); WGS (whole genome sequence); genome assembly and annotation; k-mer analysis; antimicrobial peptide; Saudi Arabia

Citation: Aljahdali, M.O.; Molla, M.H.R.; Filfilan, W.M. Whole Genome Sequence of the Newly Prescribed Subspecies *Oreochromis spilurus saudii*: A Valuable Genetic Resource for Aquaculture in Saudi Arabia. *J. Mar. Sci. Eng.* **2021**, *9*, 506. https://doi.org/10.3390/jmse9050506

Academic Editors: Nguyen Hong Nguyen and Kyall Zenger

Received: 19 April 2021
Accepted: 6 May 2021
Published: 8 May 2021

1. Introduction

Tilapias are a popular species for aquaculture around the world due to their high survival rate, fast growth, and low cost of production [1,2]. They are one of the most prominent aquaculture species and have been cultivated for 2500 years [3]. They are not only the second most important species in aquaculture but also make a huge contribution to fulfilling protein demand for human consumption [4]. Currently, tilapias are found in many regions beyond their native regions all over the world [5]. Tilapia include species of five genera, including *Tristramella*, *Oreochromis*, *Sarotherodon* [6] *Tilapia, and Danakilia* [7]. Among them, *Sarotherodon*, *Oreochromis*, and *Tilapia* are larger species as compared to others in wild fisheries as well as in aquaculture. Oreochromis mainly originated from Africa and the Middle East but are now distributed all over the world beyond their native range. Therefore, *Oreochromis* not only survive in fresh water and brackish water but also easily adapt to the high salinity of seawater due to their genetic changes. Consequently, they have the capacity to add more value to the genetic sector in the development of aquaculture [1].

Many tilapias can tolerate a wide range of salinity, from <10 to 20 ppt, and some species grow well in up to 35 ppt salinity [8,9]. The immediate ancestors of tilapia are clearly freshwater species [10]. A deep cichlid ancestor may have adapted to brackish water. The sister groups to cichlids are marine, but that goes back almost 100 million year [11]. *O. spilurus* has become a widespread aquatic animal specifically in marine ecosystems due to its adaptability and rapid growth, its ability to feed on algae and zooplankton, and

its toleration of a wide range of salinity [12]. Other tilapia species such as *O. niloticus* (Linnaeus, 1758), *O. mossambicus* (Mozambique tilapia; Peters, 1852) and *O. aureus* (blue tilapia; Steindachner, 1864) show a comparatively lower grade of saline tolerance than *O. spilurus* from the Red Sea, which is successfully cultured in marine water (salinity ~42 ppt) in Saudi Arabia [13].

The consistent production of viable seedstock is a critical objective for sustainable aquaculture production, and a whole genome sequencing study will open a new avenue for this [9,14]. The improvement of the growth rate and control of the sex ratio in tilapia production would benefit from the development of genetic markers for brood stock selection. However, tilapia culture is increasing globally due to artificial sex reversal, which has been modified into the mono-sex tilapia for large scale production [15]. Therefore, it is important to know the genetic factors of tilapia species which contribute to their consistent production. The genetic adaptions are protected from diseases which threaten all fish species due to abrupt climate change [16].

Whole genome sequencing is a new platform that helps to understand the varieties of fish disease in different traits. Therefore, it has also been used for the purposes of genetically selective breeding in fish, as well as in chromosome-level genome assembly. This next-generation sequencing technology was employed de novo in various studies for linkage mapping and quantitative traits linkage (QTL) among similar fish species [2]. The development of a genome sequence for *O. spilurus* will improve our understanding of the genetic basis for salt tolerance, growth rate, and disease resistance in this species. The earliest characterization of tilapia genomes was the construction of a genetic map for the 22 pairs of chromosomes of *O. niloticus* [1]. The initial genome sequence for tilapia was based on short Illumina sequencing reads [17]. More contiguous and complete assemblies of the tilapia genome have been retrieved from long PacBio sequencing reads [18]. Most of the cichlids are sequenced using short Illumina sequencing reads [19]. Here, we assemble a draft genome of the cultured marine tilapia *Oreochromis spilurus* using Illumina Hi-Seq sequencing technology.

2. Materials and Methods

2.1. Preparing and Sequencing Sample

The female tilapia (*O. spilurus*) subspecies was collected from seawater ponds at the Jeddah Fisheries Center on the Red Sea, Saudi Arabia. All experimental procedures were approved by the Unit of Biomedical Ethics Research Committee at King Abdul-Aziz University (Reference No. 653-20). Genomic DNA was extracted from muscle tissue using Qiagen Genomic Tip100 (Qiagen, Germantown, MD, USA). Three separate short-insert libraries and four long-insert libraries were constructed according to standard kit (Illumina, San Diego, CA, USA) protocols as outlined in the manufacturers' instructions (Figure 1). The Illumina HiSeq platform was used for the whole genome shotgun sequencing strategy, using 125 bp paired-end sequencing reads. After quality control, DNA samples were fragmented by using a covaris sonicator. Libraries were constructed by end-repair, A-tailing, adapter ligation, and size selection. The constructed library was sent for PE sequencing on an Illumina HiSeq. The pipeline is shown as follows: The picture on the left is for library preparation for small fragment sizes (e.g., 230, 350, 500 bp). The picture on the right is for library preparation for large fragment sizes (e.g., 2, 5, 10, 20 k).

2.2. Raw Sequence Data and Quality Control

The raw reads were filtered to remove duplicate reads and low-quality bases. First, we filtered reads containing adapter sequences or containing >10% of "N" base calls. Next, we evaluated the sequence quality by calculating quality scores [3]. The quality of the sequencing data is mainly distributed over Q30% (The percentage of bases with higher Ph-red score than 20 and 30 in total bases), which should ensure the integrity of downstream analyses.

2.3. Estimation of Genome Size and Assembly

The distribution of k-mer abundance was used to estimate genome size, repeat structure, and heterozygosity rate from the WGS data [4]. We used the formula G = k-mer number/k-mer depth to estimate the genome size [5]. SOAPdenovo (v2.1) software was used to assemble the sequence reads. This software first builds sequence contigs from overlaps of the individual reads. Then, it uses the paired-end sequences with variant insert size (2, 5, 10, and 20 k) to build scaffolds of the sequence contigs (Figure 2). The genome assembly (SUB6140135) and raw reads (SRR9614879) were deposited by Mohammed O. Aljahdali, King Abdulaziz University, in the NCBI database with BioProject ID: PRJNA551931. The *Oreochromis spilurus* (BioSample SAMN12169131) whole genome shotgun (WGS) project has the project accession VSJB00000000 and consists of sequences VSJB01000001–VSJB01221829. The gene accession encoded number was GCA_008269305.1.

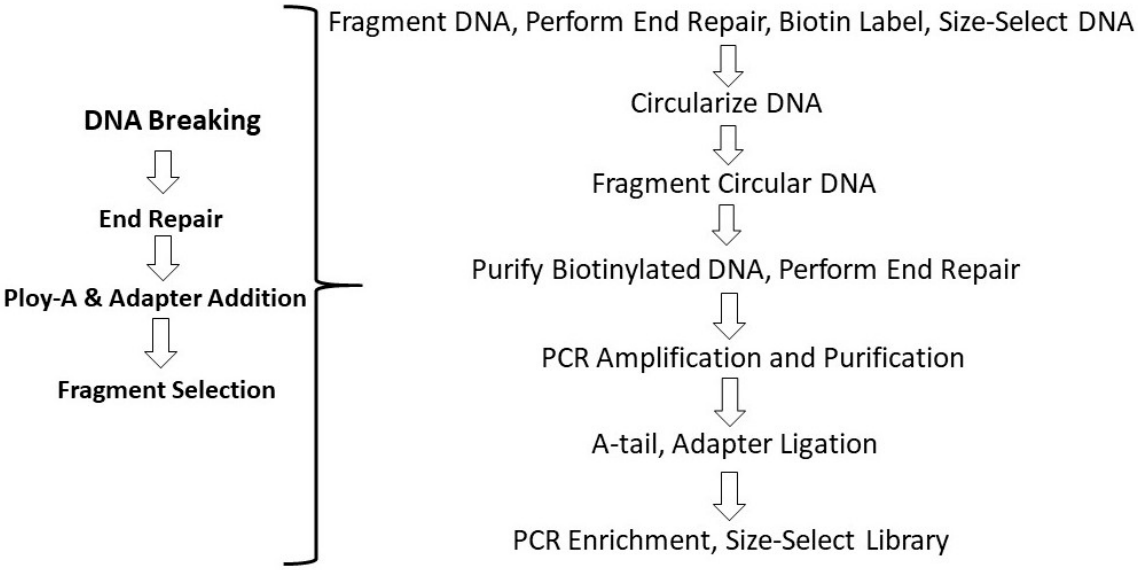


Figure 1. Workflow of library preparation and sequencing.

2.4. Phylogenetic and Divergence Time Trees

A random sampling of data from 10,000 high-quality reads were submitted to the NCBI nucleotide database for BLAST search. The top six species which have a high mapping rate were listed. One-to-one single-copy gene relatives were taken after extracting the constructed phylogenetic tree. A Perl script was used to convert their coding sequence after the protein alignment. The nucleotide sequences were linked to species by the continuous sequence. MrBayes software was used to build a phylogenetic tree, using the new sequence subsequently obtained from the BLAST [6]. The evolutionary analysis of *O. spilurus* was conducted with six other fish by MEXA X [7,8].

2.5. High-Throughput Identification of Antimicrobial Peptides

Predicted protein sequences were submitted and searched by a well-known tool (CAMPSign) to identify the antimicrobial peptide [9]. The 45 major AMP (antimicrobial peptide) families listed were searched in the CAMPR3 database based on family signature. Quarry homologous sequences were collected from the online database by using the antimicrobial peptide database (APD), CAMP, AVPdb, and BACTIBASE. The identified FASTA sequence was performed by TBLASTN (e-value (1.0×10^{-5})). The redundant results were removed after alignment hits, and an alignment indicator was used as a pairwise measurement. Moreover, hepcidin sequences of various fish were downloaded from the NCBI database, and multiple sequence alignment of hepcidin was performed by MEGA X.

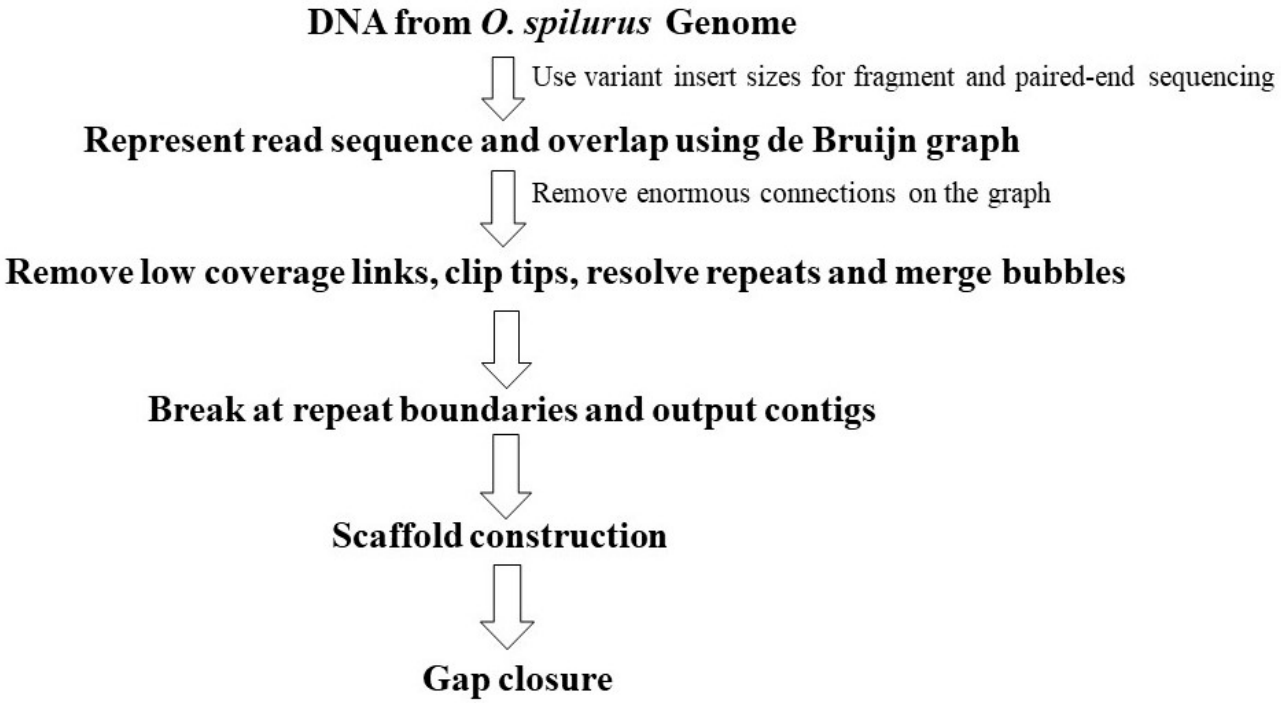


Figure 2. Assembly workflow protocol through SOAPdenovo software.

3. Results

3.1. Genome Assembly and Annotation

The low-quality reads were removed, and only clean data were obtained for subsequent genome assembly, adapter sequences, and PCR duplicates through the sequenced 205,924,310 Illumina raw reads. Concrete calculating methods were used to estimate and assemble genome size by the well-known Illumina Read Length formula: 205,924,310 × 150 bp paired-end run × 2 = 61,777,293,000 (bp). The statistical effective rate was 99.81% and the shown error rate was only 0.03%. We discarded the paired-end reads when either read contained the paired-end reads (adapter) contamination, uncertain nucleotides of more than 10 percent, and low-quality nucleotides (base quality <5) of more than 20 percent. SOAPdenovo v. 2.1 software was used for the de novo assembling of the *Oreochromis spilurus* genome [10]. The total number of scaffolds (221,829) was assembled where the N50 value contained 0.22 Mb. On the other hand, a total of 221,829 contigs were assembled where the N50 value contained 11.85 kb (Kilobase) (Table 1).

Table 1. Genome assembly and annotation of tilapia (*O. spilurus*) are estimated.

Contents	*O. spilurus*
Genome assembly	
Contig N50 size (Kb)	11.85
Scaffold N50 size (Mb)	0.22
Raw paired reads	205,924,310
Assembled genome size (Gb)	0.76
Genome annotation	-
Predicted protein-coding peptides	51,642
Predicted protein-coding genes	10,641
Conserved vertebrate genes	3144
Conserved annotated vertebrate genes with other species	2046

The popular web server GENSCAN at MIT [11] can estimate the total number of genes from the *O. spilurus* and here estimated that the number of protein-coding peptides was 51,642 (Table S1 and Table 1). Predicted protein-coding gene was produced by AUGUSTUS

(v3.3) [12] (Table S2). The conserved vertebrate gene analysis was performed using Benchmarking Universal Single-copy Orthologs (BUSCO) version v. 4.0.6. The total number of conserved vertebrate genes was 3144 (Table S3). The annotated single-copy vertebrate genes totalled 2046 (Table S4). The estimated genome size of *O. spilurus* was approximately 1.03 Gb, and the revised genome size was 1.02 Gb, which was attained by performing the routine k-mer approach. Before assembly, we estimated the genome size by k-mer analysis using a k-mer length of 17 bases. The peak of the 17 k-mer depth distribution was 53 (Figure 3), and the genome size was estimated using the formula G = N/K. The number of corresponding k-mers was 55,079,128,903. The heterozygous rate was only 0.35%, and the repeat content presented 47.97%.

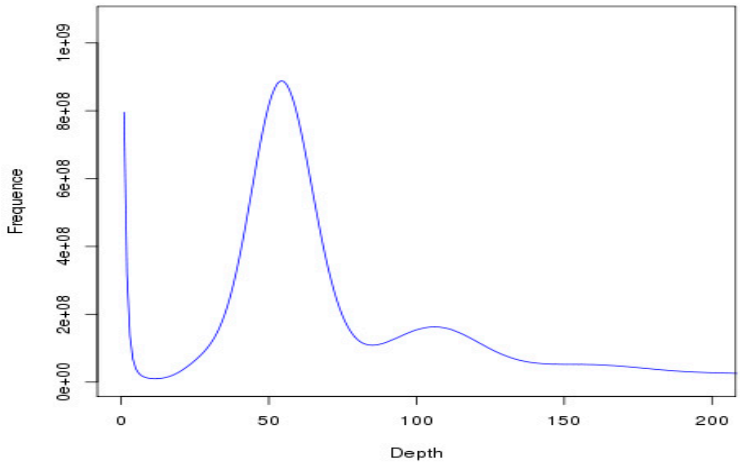


Figure 3. Frequency distribution of k-mer depth. X-coordinate is k-mer depth. Y-coordinate is the frequency of each k-mer depth.

During the validation of the assembly quality from the conserved genes, BUSCO estimated 1185 hits from complete genes. This read mainly covered the N50 scaffold length from the genome assemblies. We used the eukaryote_odb10 data set where the number of species was 70. The total number of BUSCO groups searched was 225 for inquiring the conserved gene species (Table 2). The BUSCO complete score was 173, whereas the complete and single-copy BUSCOs (S) gene family was 171 (Table S6). The total number of complete and single-copy BUSCOs (171) was twice that of the fragmented BUSCOs (86).

Table 2. Gene space completeness metrics for draft assemblies in this (BUSCO version 5.0.0) data set. (Out of 248 highly conserved eukaryotic genes).

Species	*O. spilurus*
BUSCO Complete	173
Complete and Single-copy BUSCOs (S)	171
Complete and Duplicated BUSCOs (D)	2
Fragmented BUSCOs (F)	86
Total BUSCO Groups Searched	255

The circular graph shows that the total gene size with the predicted peptide sequences(s) was 51,642. The probability of suboptimal exons was mainly determined by a cutoff value of less than 0.10. However, the total number of exons was 350,181, whereas internal exons totalled 247,753. The initial and terminal exons totaled 49,704 and 50,077, respectively (Figure 4 and Table S1).

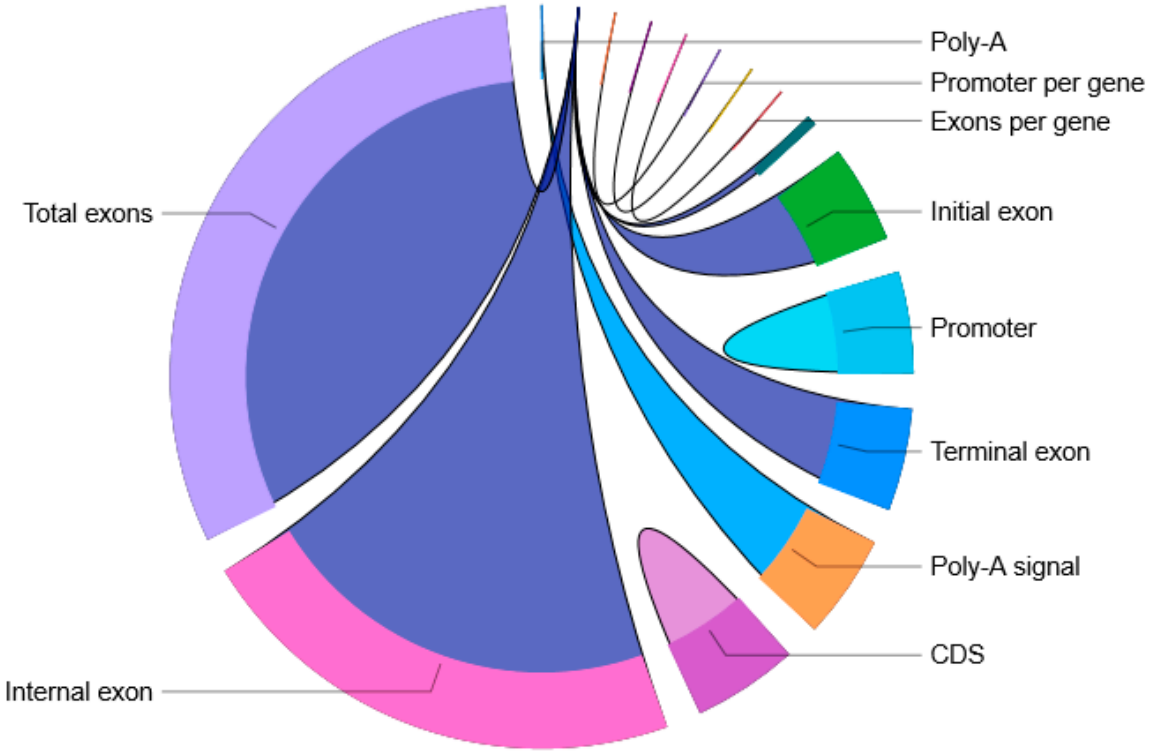


Figure 4. The circular graph depicts the gene annotation of the assembled *O. spilurus* genome.

3.2. Sequencing Depth Analysis and GC Content

GC content distribution was used to check whether there were isochores with different base compositions. Theoretically, the content of G and C, as well as A and T, should be equal in every sequencing cycle and should be stable throughout the sequencing procedure, producing a horizontal line in the results. In reality, the first several bases of the read normally display large fluctuations because of DNA amplification bias and/or low quality of the bases at the beginning, which is normal (Figure 5).

3.3. Summary of Phylogeny Tree

A random sampling of the data from 10,000 reads with higher quality was aligned with the NCBI nucleotide database (NT database) by BLAST. The top six species with high mapping rates were listed. After extracting one-to-one single-copy gene families, a phylogenetic tree was constructed. Seven protein sets of fish that included *Cyprinus carpio*, *Maylandia zebra*, *Oryzias latipes*, *Haplochromis burtoni*, *Neolamprologus brichardi*, and *Haplochromis chilotes* were downloaded from the popular Ensembl database. Molecular Evolutionary Genetics Analysis (MEGA X) was used to analyze each single-copy gene with the protein sequence of other species. The time tree shown was generated using the Realtime method. Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method and Tamura–Nei model. The divergence of *O. spilurus* from six other fish is predicted to have started form about 22.82 million years ago (Figure 6).

3.4. Antimicrobial Peptides in Oreochromis Spilurus

A total of 262 putative AMP genes were identified from the *O. spilurus* by using a local reference database (Table S5). For high–throughput identification of antimicrobial peptide, we employed BLAST to search an annotated gene set of *O. spilurus.* This covered 31 families where thaumatin was the highest in number (Figure 7). Cathelicidins are a diverse group of AMPs that are effective against bacteria, fungi, and viruses and have a

high tolerance to salt. In the study, 22 cathelicidin AMPs were documented among the 262 AMPs.

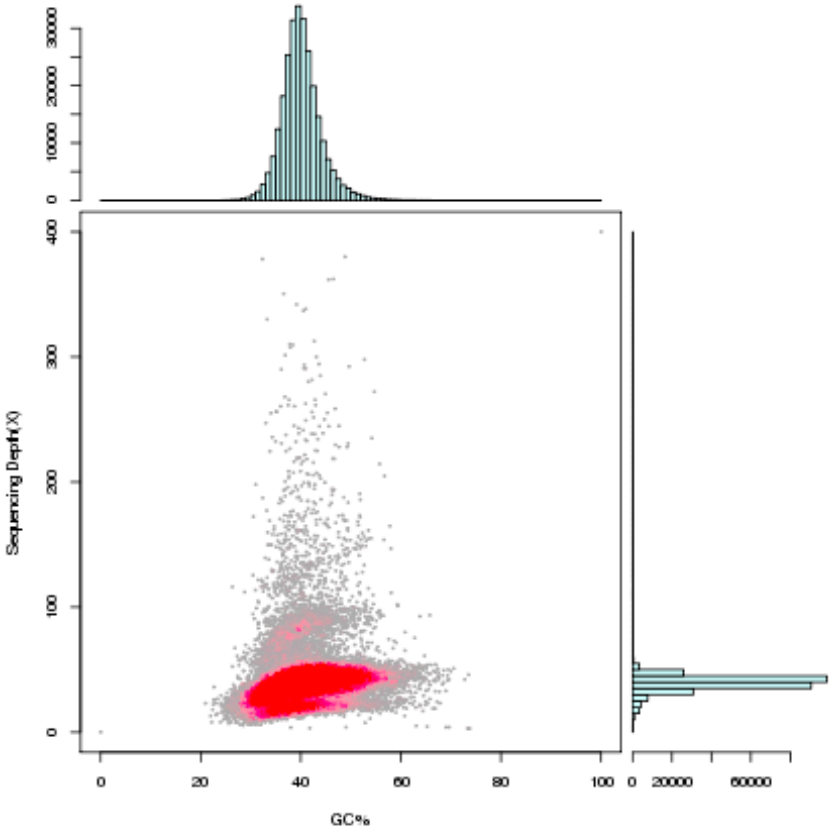


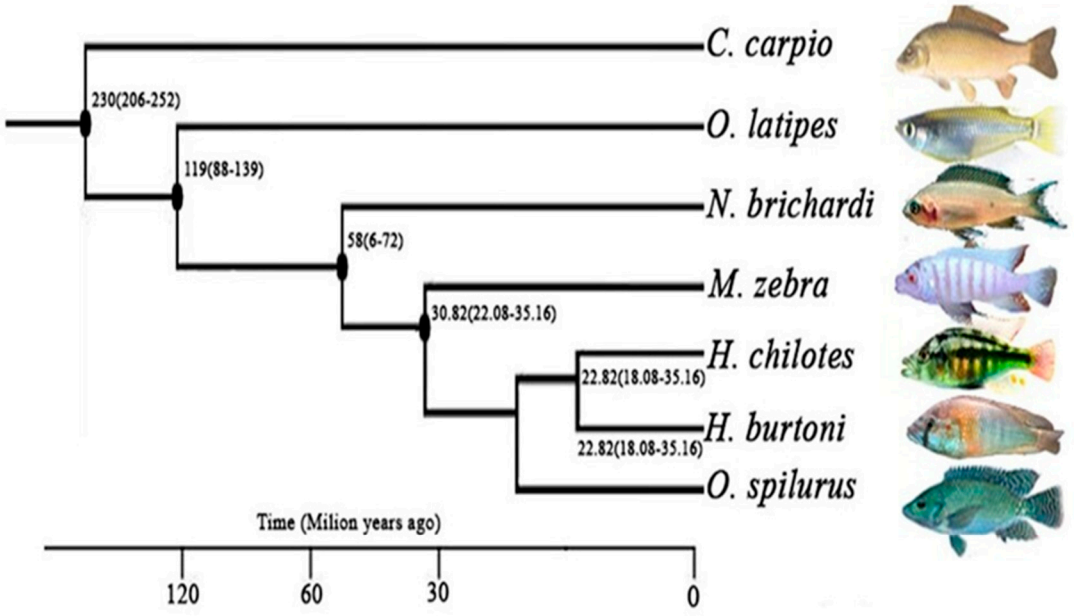
Figure 5. Correlation of GC content with sequence depth. X-coordinate is GC content, and Y-coordinate is average depth. The righthand section is the distribution of sequence depth. The upper section is the distribution of GC content. Red color represents high density where the genomic GC content was 40.4%.

GO enrichment analysis predicted that the 262 putative AMP genes from *O. spilurus* were enriched to 622 pathways through the cellular components (29%), biological process (58%), and molecular function (29%). Therefore, it interacted with different categories of acute and chronic diseases with the main classes related to immune function and diseases. The included AMP families have separate activities such as antimicrobial, antibacterial, antifungal, antiviral, and antitumor activities (Table S5 and Figure 8).

3.5. Annotation of the Putative AMP Genes

Hepcidin is an essential peptide in fish bodies partially responsible for antibacterial, antifungal, and antimicrobial activities. The number of hepcidin family AMPs in *O. spilurus* was compared with the number in other fish through multiple alignment. Hepcidin shows a >50% similarity of conserved sequences with other fishes (Figure 9A). Hepcidin (Query ID: CAMPSQ5019) has a strong antimicrobial activity as proved by the study of *Vibrio vulnificus* and *Staphylococcus aureus* from the orange-spotted grouper. Hepcidin antimicrobial peptide (Query ID: CAMPSQ2170) has a great role to play in activating innate immunity against bacteria, and similar activities are observed in bony fish studies of gilthead seabream. This hepcidin antimicrobial peptide has not been documented until now in any other *Oreochromis* spp., which have shown a >50% similarity of conserved sequences compared with other fishes (Figure 9B). Finally, a hepcidin-like antimicrobial peptide (Query ID: CAMPSQ6966) was shown to play an active role against Gram-positive bacteria (*Listeria monocytogenes* and *Enterococcus faecium*) in a study of Mozambique tilapia (*Oreochromis*

mossambicus). This putative hepcidin gene was identified by multiple sequence alignment where it showed a >80% sequence similarity compared with other fishes (Figure 9C).



Library name	The first comparison result	The second comparison result	The third comparison result	The fourth comparison result	The fifth comparsion result	The sixth comparsion result
F1	Haplochromis chilotes (14.86%)	Neolamprologus brichardi (7.89%)	Haplochromis burtoni (7.79%)	Oryzias latipes (5.23%)	Maylandia zebra (4.41%)	Cyprinus carpio (2.87%)

Figure 6. The phylogenetic position of the *O. spilurus* was determined based on one-to-one orthologues from the seven fish species, and the divergence tree was predicted with reference (black dots) to the time tree.

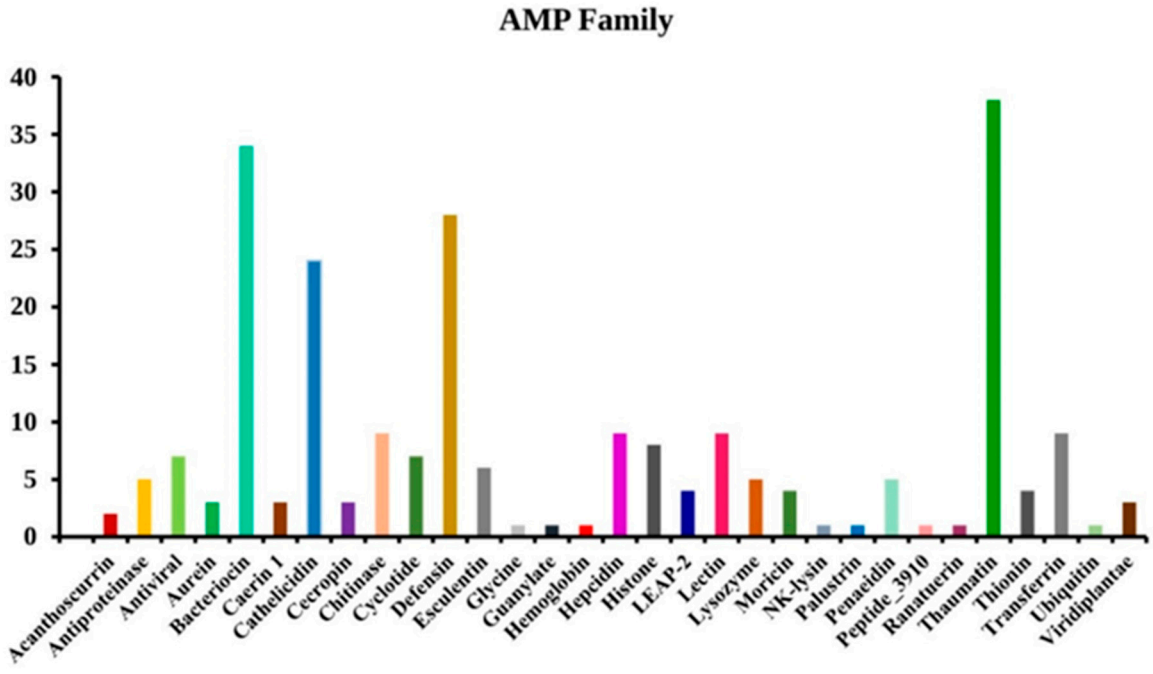


Figure 7. Records of diverse AMP families included in the *O. spilurus*.

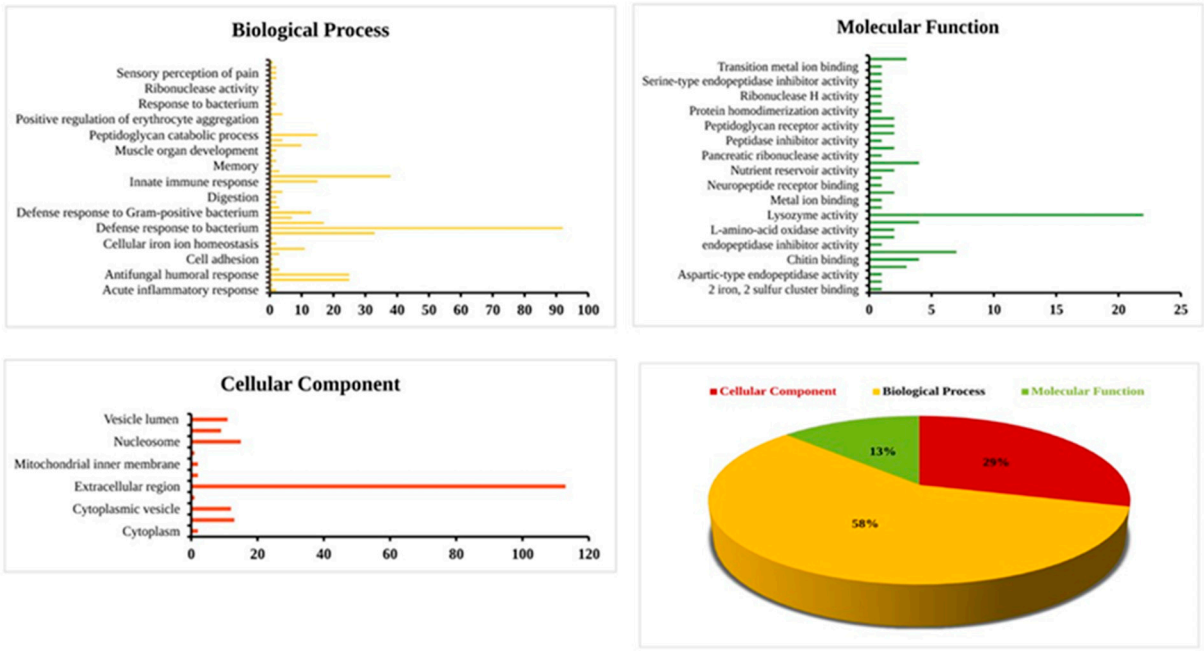


Figure 8. Annotation of the putative AMP gene from *O. spilurus*.

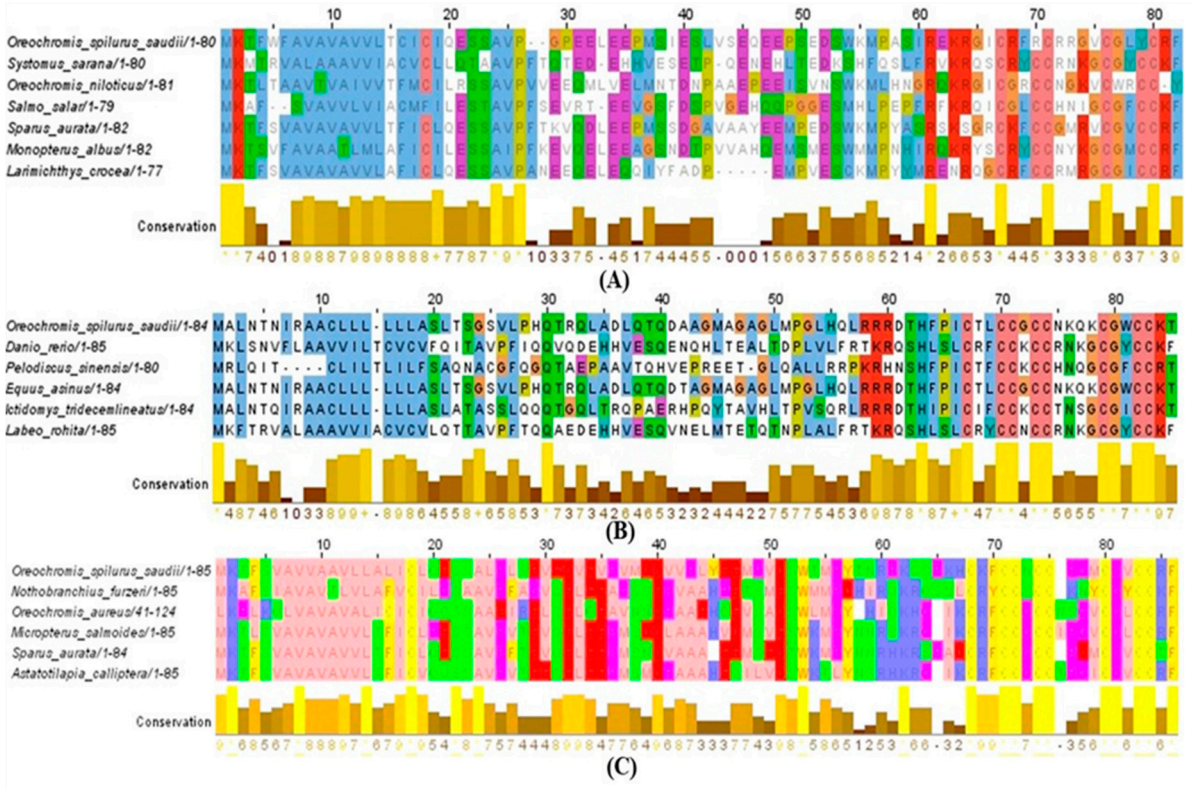


Figure 9. Putative hepcidin gene in fishes through the multiple sequence alignment. (**A**) Hepcidin; (**B**) hepcidin antimicrobial peptide; (**C**) hepcidin-like. Conserved areas representing an identity of > 50% and > 80%, respectively.

3.6. Newly prescribed Subspecies Identification

A new subspecies of marine tilapia (*Oreochromis spilurus saudii*) was discovered in seawater ponds at the Jeddah Fisheries Center on the Red Sea of Saudi Arabia using whole genome sequencing (Figure 10). They have the distinguishing characteristics from other tilapia species. High salinity adaptive capability is one of them. They can easily survive

and be cultivated in a salinity from 42 ppt to 43 ppt, better than the three other subspecies (*O. spilurus spilurus*, *O. spilurus niger*, *O. spilurus percivali*) of *O. spilurus* [13]. The sample of tilapia collected herein weighed 45 gm and had an estimated size of 8 to 9 cm. According to Jeddah Fisheries Center, the maximum weight of tilapia is 1500 gm and documented size 30 to 35 cm, which was reported from the study area. The maximum length was recorded from tilapia (26–30 cm), and the minimum length was recorded in the same species (22–23 cm). Furthermore, the growth rate is higher than other saline water species (e.g., *Oreochromis mossambicus*) in the same ponds, whose development is usually limited to 15–20 cm. Therefore, the value of these other species is low because of their shape, size, and production [14]. The mature female tilapia lays eggs 7–8 times a year with an average of 120–550 eggs per hatching time. The front color of male tilapia (*Oreochromis spilurus saudii*) is lighter than that of the female. The female tilapia's chest is bigger the male's because females usually keep fry in their mouths for nursing and protection, whereas males protect from predators. The mature male (*Oreochromis spilurus*) has similar bright-blue anal and pelvic fins to our newly identified subspecies *Oreochromis spilurus saudii.* Finally, our newly identified subspecies of tilapia can survive and be reared in 37° C [15], while *Oreochromis spilurus* are usually raised in 24–28°C [16]. Therefore, the new subspecies combine the morphological differences of the other three sub species *O. spilurus.*

Figure 10. Newly identified *O. spilurus saudii* from the Jeddah Fisheries Center at the Red Sea in Saudi Arabia.

4. Discussion

Genome sequences are important for the consistent advancement of biomedical research as well as hatchery technology with molecular breeding of tilapia. In this study, low-quality reads were removed from the total number of raw pair reads (205,924,310), and the raw base (61,777,293,000 bp) was adjusted, which ensured that data were accurate. Similarly, studies have reported that the Illumina raw reads have 239.89 gigabases and after subsequently removing low-quality data, only 161.53 Gb clean data were obtained for DNA sequence [17,18]. From the assembly, a total number of 221,829 scaffolds were taken along with the N_{50} value of 0.22 Mb as well as the contig N_{50} Value of 11.85 kb. These scaffolds (N_{50} scaffolds value and contig N_{50} value) suggested that there was a high quality of genome assembly. However, high-quality genome assembly can benefit sustainable fisheries. On the other hand, the contig N_{50} size of blue tilapia and Nile tilapia was 53.2 kb and 3.11 kb [19,20].

Moreover, marine fish peptides play a fundamental role in the survival of fish, as well as defending against enemies and helping to control osmoregulation [21]. The fish peptides exhibited large amounts of antimicrobial activities, which killed both fish and human pathogens. Currently, fish peptides are being use as for the development of therapeutic

treatment (Abuine et al., 2019). In this study, the total number of predicted peptides was 51,642, which introduces new information for *O. spilurusi.* The k-mer frequency of the read data was calculated for genomic size, where frequencies depth was 53 and the calculated k-mer numeral (N) was 55,079,128,903. Liu et al. reported the same method of showing the k-mer depth [5].

The assessment of BUSCO had a high-resolution quantification for the notation of Genome and Gen sets. The lineage dataset eukaryota_odb10 estimated a number of 70 species through the BUSCO software, using 255 BUSCO groups for genome notation. Matschiner et al. reported a similar result in their technical validation of 66 teleost species. Similarly, Burge and Karlin the genome completeness of Nile tilapia using BUSCO [22].

The GC content is an important factor in understanding the genomic function and species ecology among variable species. Furthermore, the constant changes of GC content may have ecological influences that play a pivotal role in the earth's biota. In our study, the GC content was 40.4%, which highlights the depth of analysis for exposing genomic diversity. On the other hand, different studies show that GC content above 40% results in heavy isochores [23]. Therefore, it is considered that our nucleotide composition and genomic size have proper molar ratio of GC in DNA. The genome assembly of *O. spilurus* is mentioned as a good quality example due to its high completeness and long N50 scaffold. Therefore, a diverse gene set is not only important for a high-throughput screening of AMPs but also leads to a new way of investigating disease problems in aquaculture. Bacterial and viral diseases have been an emerging threat to the tilapia industry in recent years. Therefore, it has become essential to understand AMP for the fundamental solution of diseases. In this study, long read Illumina HiSeq led to the high-quality genome sequence of *O. spilurus*. Over two hundred AMP sequences were collected through the popular CAMPSign database (http://campsign.bicnirrh.res.in/) accessed on 8 February 2021 [18]. Similarly, the blue tilapia genome revealed that it has 407 AMPs. Two groups of AMPs, such as thaumatin and bacteriocin identified the highest numbers of peptide from the group. This suggests that these AMPs may have strong effectiveness against bacterial disease. Similar studies of groups show that AMPs control microbe growth and regulate innate and adaptive immune responses [2]. These genomic resources not only expose new pathways for further molecular breeding but also provide a comprehensive cataloging of putative antimicrobial peptides of *O. spilurus*.

Furthermore, cathelicidin is a molecule which has been identified as a first line of defense against microbial invasion and high salt tolerance [24]. In our study, we documented 22 cathelicidins that may have salt tolerance ability. The peptide antimicrobial design studies showed that the antimicrobial activity of cathelicidin-RC1 was salt independent and highly stable within amphibians. Previous studies have identified the genetic effects of salinity tolerance among six tilapia varieties, such as *Oreochromis aureus* (BL), *O. mossambicus* (MO), *O. niloticus* (NI), *O. niloticus* crossbreeds (RE), Mississippi commercial strain (MC), and Florida red tilapia (FL). Nile tilapia studied with regards to changing salinity showed that four specific genes (genes beta haemoglobin, Ca^{2+} transporting plasma membrane ATPase, pro-opiomelanocortin, and beta-actin) led to changes in expression. Two significant QLT intervals (chrLG4 and chrLG18) were also identified in Nile tilapia, which are considered to aid adaptability to salinity [25].

The interpretation of genetic diversity and phylogenetic assessments were assessed by molecular study, which helps to bring consistency to the fish. The phylogenetic tree was constructed by taking a random sample of 10,000 reads data, with the higher quality aligned to the nucleotide database. We adjusted six species which have a high mapping rate with our sample species. This showed the estimated divergence time of the *O. spilurus*. Similarly, the phylogenetic tree of blue tilapia mentioned nine species based on one-to-one orthologues, and divergence trees were predicted [2]. This phylogenetic tree is important to enrich the understanding of the genes, genome, and molecular sequences, which help predict more changes in the future of *O. spilurus.*

5. Conclusions

In summary, a valuable genome assembly of the newly- cultured marine subspecies *Oreochromis spilurus* (0.76 Gb) has been accomplished for the first time, with a predicted 51,642 protein-coding peptides. These predicted peptides play a role in major cellular processes and help to identify several diseases in fish. Moreover, 262 putative AMPs were identified which may aid practical molecular breeding as well as help tackle novel bacterial and viral diseases. This subspecies can survive easily in the Red Sea in Saudi Arabia (salinity 42 ppt), which suggests that this tilapia has a high saltwater tolerance gene. In future, improving the high salinity tolerance gene will add value to saltwater fish farming as well as the sustainable development of hatchery technology.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/jmse9050506/s1. The lineage dataset is shown for the total BUSCO group search. Table S1: Gene annotation of the assembled *O. spilurus* genome and predicted protein-coding peptide. Table S2: The linage data show predicted gene sequences. Table S3: Conserved vertebrate gene. Table S4: Annotated single-copy vertebrate gene. Table S5: Predicted Antimicrobial peptides (AMPs). Table S6: Annotated single-copy gene families.

Author Contributions: M.O.A.; conceptualization, M.O.A. and M.H.R.M.; methodology, formal analysis, and writing—original draft preparation, M.O.A.; supervision, writing—review and editing and funding acquisition. W.M.F.; collecting samples. All authors have read and agreed to the published version of the manuscript.

Funding: The project awarding number was 14-ENV263-03 funded by National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia.

Institutional Review Board Statement: The study was conducted according to the approval of the Unit of Biomedical Ethics and Research Committee at King Abdulaziz University (KAU) and proceed with its guidelines (Reference No. 653-20).

Informed Consent Statement: Not applicable.

Data Availability Statement: All genomic data are uploaded into the NCBI/GENEBANK database by Mohammed O. Aljahdali, King Abdulaziz University with accession ID: PRJNA551931. The genome assembly (SUB6140135) and raw reads (SRR9614879). The *Oreochromis spilurus* (BioSample SAMN12169131) whole genome shotgun (WGS) project has the project accession VSJB00000000 and consists of sequences VSJB01000001–VSJB01221829. The gene accession encoded number was GCA_008269305.1.

Acknowledgments: The authors are grateful to the Science and Technology Unit, King Abdulaziz University, for their technical support as well as cordial cooperation. The authors are deeply grateful to Tom Kocher for his reviewing and valuable comments. Special thanks to Jeddah Fisheries Center for the sample supply.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

BUSCO	Benchmarking Universal Single-copy Orthologs
Gb	Gigabase
Mb	Megabase
Kb	Kilobase
N50	The genome (50%) is in fragments of this length
WGS	Whole Genome Sequence
AMP	Antimicrobial Peptide

References

1. Ndiwa, T.C.; Nyingi, D.W.; Claude, J.; Agnèse, J.-F. Morphological variations of wild populations of Nile tilapia (Oreochromis niloticus) living in extreme environmental conditions in the Kenyan Rift-Valley. *Environ. Boil. Fishes* **2016**, *99*, 473–485. [CrossRef]
2. Lu, G.; Luo, M. Genomes of major fishes in world fisheries and aquaculture: Status, application and perspective. *Aquac. Fish.* **2020**, *5*, 163–173. [CrossRef]
3. Conte, M.A.; Gammerdinger, W.J.; Bartie, K.L.; Penman, D.J.; Kocher, T.D. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. *BMC Genom.* **2017**, *18*, 1–19. [CrossRef] [PubMed]
4. Chen, S.; Huang, T.; Zhou, Y.; Han, Y.; Xu, M.; Gu, J. AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinform.* **2017**, *18*, 91–100. [CrossRef]
5. Liu, B.; Shi, Y.; Yuan, J.; Hu, X.; Zhang, H.; Li, N.; Li, Z.; Chen, Y.; Mu, D.; Fan, W. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* **2013**, arXiv:1308.2012.
6. McGinnis, S.; Madden, T.L. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **2004**, *32*, W20–W25. [CrossRef]
7. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]
8. Ronquist, F.; Teslenko, M.; Van Der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **2012**, *61*, 539–542. [CrossRef]
9. Waghu, F.H.; Barai, R.S.; Idicula-Thomas, S. Leveraging family-specific signatures for AMP discovery and high-throughput annotation. *Sci. Rep.* **2016**, *6*, 24684. [CrossRef] [PubMed]
10. Zhang, W.; Chen, J.; Yang, Y.; Tang, Y.; Shang, J.; Shen, B. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLoS ONE* **2011**, *6*, e17915. [CrossRef]
11. New GENSCAN Web Server at MIT. Available online: http://argonaute.mit.edu/GENSCAN.html (accessed on 13 March 2021).
12. Stanke, M.; Morgenstern, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **2005**, *33* (Suppl. 2), W465–W467. [CrossRef] [PubMed]
13. Suresh, A.V.; Lin, C. Tilapia culture in saline waters: A review. *Aquaculture* **1992**, *106*, 201–226. [CrossRef]
14. Chapman, F.A. *Culture of Hybrid Tilapia: A Reference Profile*; University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences: Gainesville, FL, USA, 2000.
15. Kosai, P.; Sathavorasmith, P.; Jiraungkoorskul, K.; Jiraungkoorskul, W. Morphometric characters of Nile tilapia (Oreochromis niloticus) in Thailand. *Walailak J. Sci. Technol.* **2014**. [CrossRef]
16. Potential New Species in the Kingdom of Saudi Arabia: Sabaki Tilapia (Oreochromis Spilurus). Available online: https://enaca.org/?id=1102&title=sabaki-tilapia-aquaculture-in-saudi-arabia (accessed on 19 February 2021).
17. Yang, Z.; Rannala, B. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol. Biol. Evol.* **2005**, *23*, 212–226. [CrossRef] [PubMed]
18. Bian, C.; Li, J.; Lin, X.; Chen, X.; Yi, Y.; You, X.; Zhang, Y.; Lv, Y.; Shi, Q. Whole Genome Sequencing of the Blue Tilapia (Oreochromis aureus) Provides a Valuable Genetic Resource for Biomedical Research on Tilapias. *Mar. Drugs* **2019**, *17*, 386. [CrossRef] [PubMed]
19. Brawand, D.; Wagner, C.E.; Li, Y.I.; Malinsky, M.; Keller, I.; Fan, S.; Simakov, O.; Ng, A.Y.; Lim, Z.W.; Bezault, E.; et al. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **2014**, *513*, 375–381. [CrossRef]
20. Symonová, R.; Suh, A. Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. *Mob. DNA* **2019**, *10*, 1–8. [CrossRef]
21. Bostock, J.; McAndrew, B.; Richards, R.; Jauncey, K.; Telfer, T.; Lorenzen, K.; Little, D.C.; Ross, L.; Handisyde, N.; Gatward, I.; et al. Aquaculture: Global status and trends. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 2897–2912. [CrossRef]
22. Burge, C.B.; Karlina, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **1997**, *268*, 78–94. [CrossRef]
23. Kastin, A.J. *Handbook of Biologically Active Peptides*, 2nd ed.; Academic Press: Cambridge, MA, USA, 2013.
24. Diamond, G.; Beckloff, N.; Weinberg, A.; Kisich, K. The Roles of Antimicrobial Peptides in Innate Host Defense. *Curr. Pharm. Des.* **2009**, *15*, 2377–2392. [CrossRef]
25. Ling, G.; Gao, J.; Zhang, S.; Xie, Z.; Wei, L.; Yu, H.; Wang, Y. Cathelicidins from the Bullfrog Rana catesbeiana Provides Novel Template for Peptide Antibiotic Design. *PLoS ONE* **2014**, *9*, e93216. [CrossRef] [PubMed]