Analysis of *Solea senegalensis* DNA Within a Specified Region

Rebecca Feeley[1], Thais Gallo Nisenbaum[2]*, Lindsay A Wilhelmus[1], Karen Zheng-Garcia[1]*
[1] Bioinformatics, Johns Hopkins University, Baltimore, MD, 21218, USA
[2] Biotechnology, Johns Hopkins University, Baltimore, MD, 21218, USA

* To whom correspondence should be addressed.
Name: Thais Gallo Nisenbaum
Email: tgallon1@jhu.edu

*Correspondence may also be addressed to.
Name: Karen Zheng-Garcia
Email: jzhengg1@jhu.edu

ABSTRACT

Solea senegalensis genome sequence between the regions 122200 to 1254000, from isolate Sse05_10M linkage group LG1, was analyzed for this project. Gene prediction tools, including Blastx, FGENSH, and GENSCAN, were used to identify potential proteins encoded by these regions. Our team successfully identified zinc finger protein PLAG1, proto-oncogene serine/threonine-protein kinase mos, and metalloreductase STEAP2. Metalloreductase STEAP2 was selected for further analysis due to its role in transmembrane transport and cellular functions. Fifteen homologous proteins encoding metalloreductase STEAP2 from different species were identified using Blastp. *Solea senegalensis* metalloreductase STEAP2 protein sequence was aligned with the fifteen identified homologs using Clustal Omega, and their molecular evolution was demonstrated using a phylogeny tree based on evolutionary divergence. Our team has shown the functional domains in the metalloreductase STEAP2 protein have been conserved across the different species.

INTRODUCTION

Solea senegalensis (NCBI Taxonomy ID 28829), commonly known as Senegalese sole, is a flatfish species. This organism lives in sea or lake bottoms up to 100 meters in depth in the Eastern Atlantic and the Mediterranean Sea (1). Although harmless to humans, *Solea senegalensis* can be parasitized by herring worms (Anisakis simplex), which can attach to the esophagus, stomach, or intestine walls when the fish is consumed raw, leading to herring worm disease (2). Additionally, *Solea senegalensis* are hosts of various types of viruses, such as the striped jack nervous necrosis virus and Viral Hemorrhagic Septicemia Virus (VHSV) (3).

The whole genome shotgun sequence of *Solea senegalensis* from isolate Sse05_10M linkage group LG1 (BioSample SAMN15430148 from *Solea senegalensis* RefSeq Raw sequence reads Bioproject PRJNA767562) is an NCBI reference sequence from a male sole, consisting of 42,924,012 base pairs. The genome assembly, IFAPA_SoseM_1(GCF_019176455.1 from the whole-genome shotgun project JAGKHQ000000000.1), was derived from the Illumina data submitted for *Solea senegalensis*, combining the alignments of *Danio rerio*, S. maximus, and S. semilaevis proteins, RNA sequence data, and ab initio gene predictions (4). The entire genome assembly consisted of 21 haploid

chromosomes, including LG1-LG21, and has provided insights into sex-associated markers and genome rearrangements in this species (4).

The specific sequence analyzed by our group (NC_058021.1) is from chromosome LG1, the largest chromosome in the organism consisting of 1,898 genes, 40.9% GC content, and coding for 3,045 proteins. LG1 is one of the two chromosomes containing canonical histone proteins H3.2 and H4 and retinoid X receptor alpha (rxra), which are involved in the cell cycle and transcription regulations (4). In addition, LG1 is also linked to lineage-specific Robertsonian fusions associated with changes in karyotype in the species (4).

The protein selected for further analysis is metalloreductase STEAP2, a protein with 521 amino acids and two notable functional regions, the COG2085 domain and the Ferric reductase domain. Metalloreductases are a family of enzymes in eukaryotes responsible for catalyzing the transmembrane transport of metals essential for cellular functions, such as iron and copper (5). By mediating these metabolic metals, some metalloreductases take part in molecular trafficking and in the regulation of cell proliferation and apoptosis (6). Many of the STEAP family members are characterized by a Ferric reductase domain, located at positions 288 to 435, in our selected protein (7). This domain is common in mammalian transmembrane proteins and necessary for ferric reductase activity, the transfer of electrons across the plasma membrane to facilitate the reduction of iron and copper (8). In humans, the STEAP2 homolog has been found to be highly-over expressed in different types of cancer (9).

MATERIAL AND METHODS

I. Whole Genome Shotgun Sequencing

We were given the region 1222001-1254000 of the *Solea senegalensis* DNA sequence for analysis. Detail of the given sequence: >NC_058021.1:1222001-1254000 *Solea senegalensis* isolate Sse05_10M linkage group LG1, IFAPA_SoseM_1, whole genome shotgun sequence

II. BLASTX

We aimed to find potential proteins that align with NC_058021.1:1222001-125400. Using Blastx, we searched the Reference Sequence (ReqSeq_proteins) database with the given query sequence. The default criteria, including BLOSUM 62 matrix and gap cost with the existence of 11 and extension of 1, were used. Furthermore, the first search was restricted to *Solea senegalensis*, as this was the organism of interest, while the second search did not have a restrictive organism criterion. The latter search aimed to find homologs in other organisms.

III. FGENESH

To detect multiple genes from different DNA strands and predict potential protein sequences, FGENESH, an ab initio gene prediction program, was used. The program utilized Hidden Markov Model (HMM) algorithm to predict gene structure by calculating the probability amino acids can be found at each position and taking into account insertions and deletions states. The gene structure and

signal detection are necessary to locate coding regions (10). This program can detect multiple genes and uses information from both a direct (+) DNA strand and a complementary (-) strand.

Using FGENESH 2.6 with *Austrofundulus limnaeus* organism genomic DNA parameter, we predicted potential proteins for NC_058021.1:1222001-125400.

IV. GENSCAN

GENSCAN is an ab initio gene prediction tool that identifies the complete structures of exons and introns using a Generalized Hidden Markov Model (GHMM) algorithm. GENSCAN takes gene density and unique CG regions into account to make predictions. This program can detect multiple genes and uses information from both DNA strands (11).

Using the web version GENSCAN 1.0 with vertebrate DNA parameters, we predicted potential proteins for NC_058021.1:1222001-125400. The suboptimal exon cut-off was set to the default value of 1.0 and the parameter matrix was set to HumanIso.smat.

V. BLASTP

To find homologous proteins to metalloreductase STEAP2 in other species, NCBI Blastp was used. Using the NCBI accession number for the metalloreductase protein in *Solea senegalensis* (XP_043891055.1) as the query sequence, the results were limited to the RefSeq_protein database. Homologs were selected from the results, including a few of the lower-scoring sequences to represent more distant phylogenetic relations. All proteins selected had an E-value of $<1x10^{-04}$.

To look for paralogs of metalloreductase STEAP2 within *Solea senegalensis*, Blastp was used once more with the NCBI accession number for the metalloreductase protein in *Solea senegalensis* (XP_043891055.1) as the query sequence. The results were generated from the non-redundant protein sequences database (nr) and limited by organism to only show results from *Solea senegalensis*.

VI. Clustal Omega

Clustal Omega was used to align the *Solea senegalensis* metalloreductase STEAP2 with the homologs found using Blastp. The FASTA formats of the sixteen protein sequences were entered into the Clustal Omega input form. Default values for all parameters were used.

VII. MEGA

To analyze molecular evolution, Molecular Evolutionary Genetics Analysis (MEGA) was used to compute pairwise distances and generation of a phylogeny tree for S*olea Senegalensis* metalloreductase STEAP2 and homologous proteins in other species. The sequences were aligned using the ClustalW method. Pairwise alignment had a gap opening penalty of 10 and a gap extension penalty of 0.10. Multiple alignments had a gap opening penalty of 10 and a gap extension penalty of 0.20. Pairwise distances and an unweighted pair group method (UPGMA) phylogeny tree were generated utilizing the Poisson model.

RESULTS

I. Blastx Query Resulted in Three Perfect Protein Hits

The first search, which limited the organism hits to *Solea senegalensis*, resulted in one hundred potential proteins. Three proteins had 100 percent identity, and 0.0 or $1x10^{-109}$ E value, indicating near-perfect, if not perfect, homology between the query protein and the potential proteins. One uncharacterized protein with gene symbol LOC12278453 and E value of $2x10^{-17}$ also resulted in a homologous match.

The potential proteins with given names are zinc finger protein PLAG1, proto-oncogene serine/threonine-protein kinase mos, and metalloreductase STEAP2. Zinc finger protein PLAG1 has coding sequence (CDS) locations between 1230637 and 1232246 and shares orthologs with 363 organisms, including but not limited to, *Homo sapiens*, *Mus musculus*, *Bos taurus*, and *Gallus gallus*. Proto-oncogene serine/threonine-protein kinase mos have CDS locations between 1243181 and 1244191 and share orthologs with 360 organisms, such as *Danio rerio*, *Sus scrofa*, *Macaca mulatta*, and *Xenopus tropicalis*. Metalloreductase STEAP2 has CDS locations between 1248717-1253188 and shares orthologs with 341 organisms, such as *Pan troglodytes*, *Canis lupus familiaris*, *Cricetulus griseus*, and *Equus caballus*.

II. Comparison of Gene Prediction Outputs with Blastx Proteins

Gene prediction tools attempt to locate coding sequences of genes by identifying internal, terminal, and any existing exons. Homology-based gene prediction programs such as Blastx, FGENESH, and GENSCAN use mRNA data to identify gene expression. As mentioned above, Blastx identified three proteins with 100 percent identity to the *Solea senegalensis* DNA sequence: zinc finger protein PLAG1, proto-oncogene serine/threonine-protein kinase mos, and metalloreductase STEAP2. FGENESH 2.6 with *Austrofundulus limnaeus* genomic DNA parameter predicted six genes, four in the + chain and two in the – chain, as well as seventeen exons, ten in the + chain, and seven in the – chain. The genes and exons positions received a score of 465.70068. Genscan with the vertebrate genomic DNA parameter predicted four peptide sequences. The additional gene prediction tools successfully identified zinc finger protein PLAG1, proto-oncogene serine/threonine-protein kinase mos, and metalloreductase STEAP2 as potential proteins encoded by the *Solea senegalensis* DNA sequence.

a. Zinc Finger Protein PLAG1 Predictions

Zinc finger protein PLAG1 is 478 amino acids in length and is encoded by the pleiomorphic adenoma gene 1 (PLAG1). The corresponding mRNA sequence for the *Solea senegalensis* pleiomorphic adenoma gene 1 identified in the NCBI nucleotide database has CDS locations between positions 237 and 1673, and a polyA site at position 4717. Using the NCBI gene database, the mRNA for *Solea senegalensis* PLAG1 gene contains three exons at locations between positions 1 to 127, 6162 to 6518, and 6692 to 7880. The CDS locations are between positions 6162 to 6518 and 6692 to 7880, which means one of the exons is in the untranslated region of the mRNA and would not be detected by the gene prediction tools.

FGENESH: The protein sequence predicted by FGENESH is identical to the one identified in Blastx in terms of both length and amino sequence. FGNESH predicted a transcription start site at position 7906, CDS locations between positions 8637 and 10246, a polyA site at position 10433, and two exons at positions 8637 to 8884 and 9058 to 10246.

GENSCAN: GENSCAN predicted a sequence of 548 amino acids in length, but 478 of them match the other two predicted proteins exactly, as confirmed by Blastp. four exons were predicted at positions 6547 to 6600, 7149 to 2399, 8712 to 8964, and 9138 to 10326, and a polyA site between positions 10932 to 10937.

b. Proto-oncogene Serine/Threonine-Protein Kinase Mos Predictions

Proto-oncogene Serine/Threonine-Protein Kinase Mos is 336 amino acids in length and is encoded by the v-mos Moloney murine sarcoma viral oncogene homolog (mos). The corresponding mRNA sequence for *Solea senegalensis* v-mos Moloney murine sarcoma viral oncogene homolog (mos) has CDS locations between positions 254 and 1264 and a polyA site at position 1799. Using the NCBI gene database, the mRNA for *Solea senegalensis* mos gene contains one exon at locations between positions 1242928 to 1244726 with a CDS region at locations between positions 243181 to 1244191.

FGENESH: The protein sequence predicted by FGENESH is identical to the one identified in Blastx in terms of both length and amino sequence. FGNESH predicted a transcription start site at position 20545, CDS locations between positions 21181 and 22191, a polyA site at position 23043, and one exon at positions 21181 to 22191.

GENSCAN: GENSCAN predicted a sequence of 453 amino acids in length, but 336 of them match the other two predicted proteins exactly, as confirmed by Blastp. Three exons were predicted at positions 221261 to 22267, 22334 to 22497, and 24780 to 24970, and a polyA site between positions 26554 to 26559.

c. Metalloreductase STEAP2 Predictions

Metalloreductase STEAP2 is 521 amino acids in length and is encoded by the STEAP family member 2, metalloreductase (STEAP2). The corresponding mRNA sequence for *Solea senegalensis* STEAP family member 2, metalloreductase (steap2) CDS locations between positions 193 and 1758, and a polyA site at position 4342. Using the NCBI gene database, the complementary mRNA for *Solea senegalensis* for the STEAP2 gene contains 6 exons at locations between positions 1246133 to 1249010, 1249751 to 1249915, 1250244 to 1250771, 1252510 to 1252676, 1252777 to 1253238, 1254336 to 1254477. CDS locations are between positions 1248717 to 1249010, 1249751 to 1249915, 1250244 to 1250771, 1252510 to 1252676, and 1252777 to 1253188

FGENESH: The protein sequence predicted by FGENESH is identical to the one identified in Blastx in terms of both length and amino sequence. FGNESH predicted a transcription start site at position 31588, CDS locations between positions 26717 and 31188, a polyA site at position 26578, and five exons at positions 26717 to 27010, 27751 to 27915, 28244 to 28771, 30510 to 30676, and 30777 to 31188.

GENSCAN: GENSCAN predicted a sequence of 652 amino acids in length. According to Blastp, ranges 384 to 555 are 100% identical to the other two predicted sequences and ranges 1 to 164 and 552 to 652 are 97% identical to the other two predicted sequences. Four exons were predicted at positions 27090 to 26797, 28837 to 28324, 30756 to 29913, and 31163 to 30857, and a polyA region between positions 2663 to 26658.

NCBI Data (Nucleotide Database NC_058021.1)	NCBI blastx (RefSeq database)		FGENESH 2.6 (Austrofundulus limnaeus)		GENSCAN 1.0 (Vertebrate)	
Gene: PLAG1 **mRNA:** 1-127, 6162-6518, 6692-10924) **CDS:** 6271-6518, 6692-7880 **Product:** Zinc finger protein PLAG1	>XP_043901516.1 zinc finger protein PLAG1 [Solea senegalensis] 478 aa MATGTQGHRDHILDKAKFTAPTGRRRRAEGKPKKNFPCQECQKAFNSLEKLKVHSYSHTGERPYRCSHPDCTKAFVSKYKLLRHMATHSPEKNHKCSYCEKMFHRKDHLKNHLHTHDPYKEAFTCQECGKSYNTKLGFKRHLALHAANNGDLTCQVCLQPFPSTGILLEHLKTHAGKSSGGTKEKKHRCEHCERRFYTRKDVRRHMVVHTGRKDFLCQYCAQRFGRKDHLTRHMKKSHPQELLRVKTEPADSLEPIVYDLASGDIKGELAGMLTPRPHQLNMYAGPVLDAEPFPAPTHPFSLKYPLGPNLTSYAVSSQEREQNLKGELETYLMELQNSMPSSSSAAQEPQIPSSKLELEPPGGGLEGASEDVSMSKMSTSVATPTAGDSLASSSSLMDFSQLFNFLPLNGPPYNQAGGSGGQSIAYQPSEESAPLVQLPPQSSEGPEGAESPLQGLPSSFISNLSTPTTLPRFHQAFQ	**Number of Exons Predicted:** 3 **CDS region:** 237 - 1673 **PolyA site:** 4717 **Exon 1:** 128-485 **Exon 2:** 485 - 4717 **mRNA Length:** 4717 bp **Exon 1:** 1224367-1224493 (UTR) **Exon 2:** 1230528-1230884 **Exon 3:** 1231058-1235290	>FGENESH: 3 2 exon (s) 8637 - 10246 478 aa, chain + MATGTQGHRDHILDKAKFTAPTGRRRRAEGKPKKNFPCQECQKAFNSLEKLKVHSYSHTGERPYRCSHPDCTKAFVSKYKLLRHMATHSPEKNHKCSYCEKMFHRKDHLKNHLHTHDPYKEAFTCQECGKSYNTKLGFKRHLALHAANNGDLTCQVCLQPFPSTGILLEHLKTHAGKSSGGTKEKKHRCEHCERRFYTRKDVRRHMVVHTGRKDFLCQYCAQRFGRKDHLTRHMKKSHPQELLRVKTEPADSLEPIVYDLASGDIKGELAGMLTPRPHQLNMYAGPVLDAEPFPAPTHPFSLKYPLGPNLTSYAVSSQEREQNLKGELETYLMELQNSMPSSSSAAQEPQIPSSKLELEPPGGGLEGASEDVSMSKMSTSVATPTAGDSLASSSSLMDFSQLFNFLPLNGPPYNQAGGSGGQSIAYQPSEESAPLVQLPPQSSEGPEGAESPLQGLPSSFISNLSTPTTLPRFHQAFQ	**Number of Exons Predicted:** 2 **CDS Region:** 8637 - 10246 **Transcription Start Site:** 7906 **Exon 1:** 8637 - 8884 **Exon 2:** 9058 - 10246 **PolyA Site:** 10433 **mRNA Length:** 1437 bp	>/tmp/10_12_22-18:23:21.fasta|GENSCAN_predicted_peptide_1| 548_aa MVVIVSDLFPVVLPVTQQSNHCERFEIVCRSKGKQREKLWASFTIRRGELSSHLPGSMKLVLSPHQGAMATGTCQHRDHILDKAKFTAPTGRRRRAEGKPKKNFPCQECQKAFNSLEKLKVHSYSHTGERPYRCSHPDCTKAFVSKYKLLRHMATHSPEKNHKCSYCEKMFHRKDHLKNHLHTHDPYKEAFTCQECGKSYNTKLGFKRHLALHAANNGDLTCQVCLQPFPSTGILLEHLKTHAGKSSGGTKEKKHRCEHCERRFYTRKDVRRHMVVHTGRKDFLCQYCAQRFGRKDHLTRHMKKSHPQELLRVKTEPADSLEPIVYDLASGDIKGELAGMLTPRPHQLNMYAGPVLDAEPFPAPTHPFSLKYPLGPNLTSYAVSSQEREQNLKGELETYLMELQNSMPSSSSAAQEPQIPSSKLELEPPGGGLEGASEDVSMSKMSTSVATPTAGDSLASSSSLMDFSQLFNFLPLNGPPYNQAGGSGGQSIAYQPSEESAPLVQLPPQSSEGPEGAESPLQGLPSSFISNLSTPTTLPRFHQAFQ	**Number of Exons Predicted:** 4 **Exon 1:** 6547 - 6600 **Exon 2:** 7149 - 2399 **Exon 3:** 8712 - 8984 **Exon 4:** 9138 - 10326 **PolyA:** 10932 - 10937 **mRNA Length:** 1647 bp
Gene: mos **mRNA:** 1242928 - 1244726 **CDS:** 1243181 - 1244191 **Product:** Proto-oncogene serine/threonine-protein kinase mos	>XP_043891069.1 proto-oncogene serine/threonine-protein kinase mos [Solea senegalensis] 336 aa MPSPIPVTRLLPRDIYPSLDTRACSSPLTKHGSTLQVPTERLHGERVASRLWSSVIYWKELRSLQPVGSGGFGSVYRAEYLGQTAALKRVKKCTKNRLASRQSFWAELNAAHLRHGNIVRVIAATTCVPADFGDDDCIGMILMEYVGDRNLQQVIYDSAELQEEQWLRYSTDIAHGLRFLHAHSVVHLDIKPANVLVTSDDVCKIVDFGCSFKLERGCEVSVIGPELSHVGGTYTHRAPELLKGEAVSPKADIFSFGITLWQLMTREQPYTGDRQHVVYAVVAQNLRPPVHESPVFRSEQGRLCRSLLSRCWSGDARCRPSAQELLQQLEQLRSHA	**Number of Exons Predicted:** 1 **CDS region:** 254 - 1264 **PolyA site:** 1799 **Exon 1:** 1-1799 **mRNA Length:** 1799 bp **Exon 1:** 1242928-1244726	>FGENESH: 5 1 exon (s) 21181 - 22191 336 aa, chain + MPSPIPVTRLLPRDIYPSLDTRACSSPLTKHGSTLQVPTERLHGERVASRLWSSVIYWKELRSLQPVGSGGFGSVYRAEYLGQTAALKRVKKCTKNRLASRQSFWAELNAAHLRHGNIVRVIAATTCVPADFGDDDCIGMILMEYVGDRNLQQVIYDSAELQEEQWLRYSTDIAHGLRFLHAHSVVHLDIKPANVLVTSDDVCKIVDFGCSFKLERGCEVSVIGPELSHVGGTYTHRAPELLKGEAVSPKADIFSFGITLWQLMTREQPYTGDRQHVVYAVVAQNLRPPVHESPVFRSEQGRLCRSLLSRCWSGDARCRPSAQELLQQLEQLRSHA	**Number of Exons Predicted:** 1 **CDS Region:** 21181 - 22191 **Transcription Start Site:** 20545 **Exon 1:** 21181 - 22191 **PolyA Site:** 23043 **mRNA Length:** 1011 bp	>/tmp/10_12_22-18:23:21.fasta|GENSCAN_predicted_peptide_3| 453_aa MPSPIPVTRLLPRDIYPSLDTRACSSPLTKHGSTLQVPTERLHGERVASRLWSSVIYWKELRSLQPVGSGGFGSVYRAEYLGQTAALKRVKKCTKNRLASRQSFWAELNAAHLRHGNIVRVIAATTCVPADFGDDDCIGMILMEYVGDRNLQQVIYDSAELQEEQWLRYSTDIAHGLRFLHAHSVVHLDIKPANVLVTSDDVCKIVDFGCSFKLERGCEVSVIGPELSHVGGTYTHRAPELLKGEAVSPKADIFSFGITLWQLMTREQPYTGDRQHVVYAVVAQNLRPPVHESPVFRSEQGRLCRSLLSRCWSGDARCRPSAQELLQQLEQLRSHAVTQPKTRAQFDSVSRCHGAVPCDTVKSAWRHRSRGLSSRIRLQCPVNGIVCISGSMFSVRVSVPGSGMGTTALTMYNYNHTSEKPEPSLPGRKVSKVLLLVTLIKSESSSDCSVSVSL	**Number of Exons Predicted:** 3 **Exon 1:** 221261 - 22267 **Exon 2:** 22334 - 22497 **Exon 3:** 24780 - 24970 **PolyA:** 26554 - 26559 **mRNA Length:** 1362 bp
Gene: steap2 **mRNA Complement:** 1246133 - 1249010,1249751 - 1249915, 1250244 - 1250771,1252510 - 1252676, 1252777 - 1253238, 1254336 - 1254477 **CDS Complement:** 1248717 - 1249010,1249751 - 1249915, 1250244 - 1250771, 1252510 - 1252676, 1252777 - 1253188 **Product:** Metalloreductase STEAP2	>XP_043891055.1 metalloreductase STEAP2 [Solea senegalensis] 521 aa MDSISMMSGSRSHVSLTASSSCFSSSSHHHFHHTQACFLPSCMRNGVADTGAACLAAQPTVAILGSGDFSKCLTIRLLRHGFHVVVGSRHPKQAAQSFPHVVDVTHHEDAVGKAAIVFLAIRREHYSVLWDLKHLLEGKILVDVSNNRRLNQYPESNAEYLASLLPDSVVVKGFNVISAWAMQQSCPKDASAQVFVCSDSLEARQQIMELARQLNFLPVDMGTLSSSRDIENLPIQLFSSWRGPVLAAVALSIFFFAYSFVRDVIHPYLKYKKSDFYKIPIEMVNRTLPTVAITLLALVYLAGQLAAAHQLYYGTKYRQFPRWLEGWLQSRKQMGLLSFFLAAVHVLYSLCLPMRRSERYLLLNSAYQQVHANVENSWNEEEVWRVEMYISFGIMSLGLLSLAITSIPSVHSQLNWREFSFIQSTLGYIALLIATLHGLLFGWKRAFEEESYRFYLPPSFVVALALPVCVILGKVVMLLPCIAQKLHRIRRGMDSRQYQCRRLEPVGSAAHVSPERVTIM	**Number of Exons Predicted:** 6 **CDS region:** 193 - 1758 **PolyA site:** 4342 **Exon 6:** 1465 - 4342 **Exon 5:** 1300 - 1464 **Exon 4:** 772 - 1299 **Exon 3:** 605 - 771 **Exon 2:** 143 - 604 **mRNA Length:** 1799 bp **Exon 6:** 1246133-1249010 **Exon 5:** 1249751-1249915 **Exon 4:** 1250244-1250771 **Exon 3:** 1252510-1252676 **Exon 2:** 1252777-1253238 **Exon 1:** 1254336-1254477	>FGENESH: 6 5 exon (s) 26717 - 31188 521 aa, chain - MDSISMMSGSRSHVSLTASSSCFSSSSHHHFHHTQACFLPSCMRNGVADTGAACLAAQPTVAILGSGDFSKCLTIRLLRHGFHVVVGSRHPKQAAQSFPHVVDVTHHEDAVGKAAIVFLAIRREHYSVLWDLKHLLEGKILVDVSNNRRLNQYPESNAEYLASLLPDSVVVKGFNVISAWAMQQSCPKDASAQVFVCSDSLEARQQIMELARQLNFLPVDMGTLSSSRDIENLPIQLFSSWRGPVLAAVALSIFFFAYSFVRDVIHPYLKYKKSDFYKIPIEMVNRTLPTVAITLLALVYLAGQLAAAHQLYYGTKYRQFPRWLEGWLQSRKQMGLLSFFLAAVHVLYSLCLPMRRSERYLLLNSAYQQVHANVENSWNEEEVWRVEMYISFGIMSLGLLSLAITSIPSVHSQLNWREFSFIQSTLGYIALLIATLHGLLFGWKRAFEEESYRFYLPPSFVVALALPVCVILGKVVMLLPCIAQKLHRIRRGMDSRQYQCRRLEPVGSAAHVSPERVTIM	**Number of Exons Predicted:** 5 **CDS Region:** 26717 - 31188 **PolyA Site:** 26578 **Exon 5:** 26717 - 27010 **Exon 4:** 27751 - 27915 **Exon 3:** 28244 - 28771 **Exon 2:** 30510 - 30676 **Exon 1:** 30777 - 31188 **Transcription Start Site:** 31588 **mRNA Length:** 1566 bp	>/tmp/10_12_22-18:23:21.fasta|GENSCAN_predicted_peptide_4| 652_aa ACFLPSCMRNGVADTGAACLAAQPTVAILGSGDFSKCLTIRLLRHGFHVVVGSRHPKQAAQSFPHVVDVTHHEDAVGKAAIVFLAIRREHYSVLWDLKHLLEGKILVDVSNNRRLNQYPESNAEYLASLLPDSVVVKGFNVISAWAMQQSCPKDASAQVPTSTNIHQHPPTSTSIHQHPPTTNIHQHPPTSTSIHQHPPTSTSIHQHQQPPTTTNTHQHPPTSTSIHQHPPTTTNTHQHPPTSTSIHQHPPTTTNTHQHPPTSTSIHQHPPTTTNTHQHPPTSTSIHQHPPTSTNIHQHPPTSNNHQHPPASTSIHQHPPTSTSIHQHPPASTNIQAASTNIHQHPPTSTSQHPPTSTNIHQHPPTSTNIHQHPPASTNIHSSTLTTTNLPQSDSLEARQQIMELARQLNFLPVDMGTLSSSRDIENLPIQLFSSWRGPVLAAVALSIFFFAYSFVRDVIHPYLKYKKSDFYKIPIEMVNRTLPTVAITLLALVYLAGQLAAAHQLYYGTKYRQFPRWLEGWLQSRKQMGLLSFFLAAVHVLYSLCLPMRRSERYLLLNSAYQQSTLGYIALLIATLHGLLFGWKRAFEEESYRFYLPPSFVVALALPVCVILGKVVMLLPCIAQKLHRIRRGMDSRQYQCRRLEPVGSAAHVSPERVTIM	**Number of Exons Predicted:** 4 **PolyA:** 2663 - 26658 **Exon 4:** 27090 - 26797 **Exon 3:** 28837 - 28324 **Exon 2:** 30756 - 29913 **Exon 1:** 31163 - 30857 **mRNA Length:** 1959 bp

Table 1

III. Homologous Results for Metalloreductase STEAP2

Most of the homologs found by Blastp were other fish species. The best scoring non-predictive match was to metalloreductase in Micropterus salmoides, Large Mouth Bass. With an E-value of 0 and a percent identity of 89.25%, it is almost an identical protein to the metalloreductase in *Solea senegalensis*. The lowest scoring match included in our phylogeny is the metalloreductase found in Papio Anubis, the Olive Baboon, with 66.79% identity. Due to the importance of metalloreductases in eukaryotic cellular processes, it is not surprising to see it well conserved among many different species so that it may carry out its function.

The full list of homologous organisms include: *Micropterus salmoides* (XP_038579507.1), *Morone saxatilis* (XP_035509014.1), *Gadus morhua* (XP_030204512.1), *Thunnus albacares* (XP_044196375.1), *Mugil cephalus* (XP_047424079.1), *Oncorhynchus mykiss* (XP_021476238.1), *Xiphias gladius* (XP_039982848.1), *Chelmon rostratus* (XP_041817593.1), *Echeneis naucrates* (XP_029385537.1), *Anabas testudineus* (XP_026219050.1), *Spheniscus magellanicus* (KAF1422982.1), *Perca fluviatilis* (KAF1372749.1), *Atractosteus spatula* (MBN3317059.1), *Bos taurus* (NP_001071315.1), *Papio Anubis* (NP_001162460.1).

The *Solea senegalensis* paralogs found include products from the STEAP3 (KAG7503596.1) and STEAP4 (XP_043890536.1) genes located on chromosomes LG2 and LG9, respectively. The STEAP

proteins are all metalloreductases made in different parts of the genome, but due to their shared function, have similar amino acid sequences.

IV. Multiple Sequence Alignment of Metalloreductase STEAP2

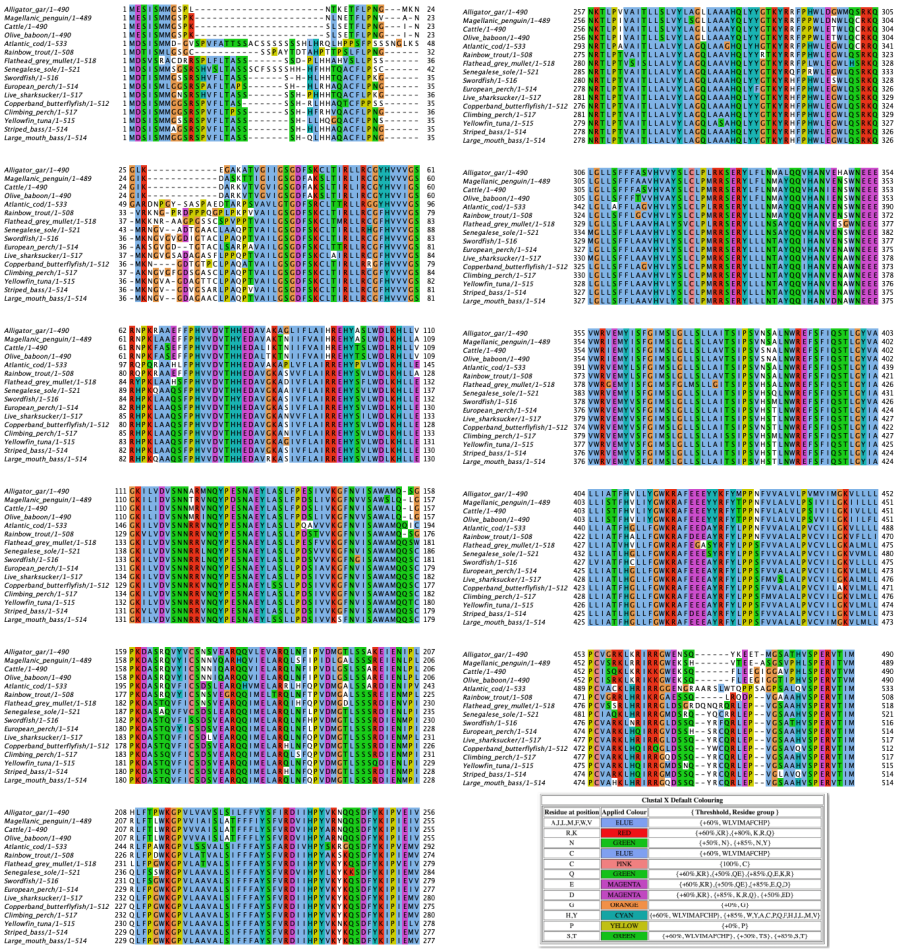

Figure 1 (12)

Multiple sequence alignment (MSA) was calculated by Clustal Omega with protein sequence type and default settings. Like the earlier Clustal W, Clustal Omega uses a progressive multiple alignment strategy (13) An exact MSA has an algorithmic complexity of O (LN) where L is the sequence length and N is the number of sequences. The time required for a solution becomes impractical once more than two sequences are being compared. Progressive alignment strategies create a guide tree, which starts with the alignment of the most closely related sequences and continues until the least closely related sequence is aligned. Earlier progressive alignments calculate all possible pairwise alignments and have an algorithmic complexity of $O(N^2)$. Clustal Omega uses mBed to produce guide trees with a complexity of O(N log N). It is a highly accurate MSA tool that works quickly on sequences of any size (13).

The results of the MSA are shown in a table generated by Jalview (14). Jalview is free software that allows customizable viewing and formatting of MSAs (15). The color scheme used is the Clustal X color scheme. Amino acids are assigned a color based on their chemistry: blue for hydrophobic amino acids, green for polar uncharged, etc (12). In special cases, glycine, proline, and cysteine get their

own colors (orange, yellow, and pink, respectively). Not all amino acids in the MSA get a color, however. Only amino acids that reach a certain threshold in their column are colored (12). The colors and the thresholds are shown in Figure 1. As an example, in position 6, 14 out of 15 amino acids are M (methionine), colored blue. However, the remaining amino acid is R (arginine). The M's are colored because 14/15 (93%) is above the threshold of 60% for the amino acids that share a row with methionine (A, I, L, F, W, and V). However, arginine makes up 1/15 (7%) of the amino acids at position 6, which is below the threshold for its row (60%). The classification of amino acid chemistries by color, as well as the use of color only when the column contains a certain threshold from that grouping, allows a quick determination of the properties of a protein at a given location.

V. Phylogeny of *Solea Senegalensis* for Metalloreductase STEAP2

The UPGMA phylogeny tree was generated based on the distance matrix, which analyzes the number of substitutions present between sequences. The UPGMA tree is rooted, which specifies an evolutionary path.

As depicted by the tree, the homolog closest related to *Solea senegalensis* metalloreductase STEAP2 is the *Micropterus salmoides*, with a pairwise distance value of 0.1023, followed by *Anabas testudineus* with a pairwise distance value of 0.1043. The furthest homologs are the *Papio Anubis* and *Atractosteus spatula*, both containing a pairwise distance value of 0.3439, followed by *Bos taurus*, with a pairwise distance value of 0.3324.

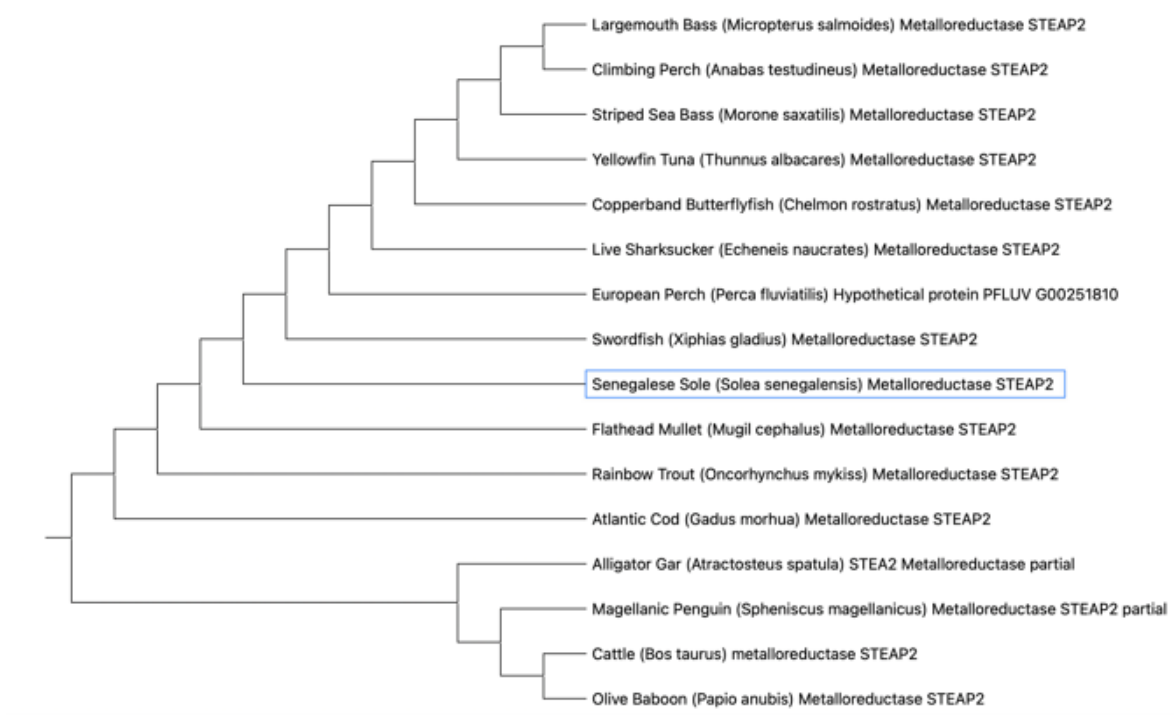

Figure 2

DISCUSSION

As a result of our investigations on a section of chromosome LG1 in *Solea senegalensis*, our team was able to discover the protein-coding genes in the sequence and their products, as well as using homology and phylogeny to investigate evolutionary relationships to our organism of interest.

In order to do this, our team utilized a wide variety of web-based tools and algorithms, each with its own strengths and weaknesses. Using NCBI's Blastx, our team has identified zinc finger protein PLAG1, proto-oncogene serine/threonine-protein kinase mos, and metalloreductase STEAP2 as potential proteins coded for between the regions 1222001 to 1254000 of this sequence. Additional gene-prediction tools, such as FGENESH and GENSCAN, were used to compare the predicted protein sequences. Using NCBI's Blastp tool to gather homologous proteins for metalloreductase gave us reliable results for use in this paper, however, using PSI-Blast for finding an even more distantly related species may have been an alternative option for investigating phylogeny or a more in-depth analysis of distant evolutionary relationships. The software MEGA was used to generate a UPGMA phylogeny tree, demonstrating molecular evolution and divergence between the organisms.

Knowing the importance of the family of STEAP proteins in eukaryotic cellular functions, our team chose metalloreductase STEAP2 in *Solea senegalensis* as our protein of interest for the homology and phylogeny research. The results supported the hypothesis that there would be many homologous proteins in a variety of species as sequence and function are closely related in proteins, functional domains would be well conserved.

DATA AVAILABILITY

Whole genome shotgun sequence for *Solea senegalensis* isolate Sse05_10M linkage group LG1, IFAPA_SoseM_1 found in NCBI Nucleotide database (NC_058021.1).

Protein sequences for zinc finger protein PLAG1, proto-oncogene serine/threonine-protein kinase mos, and metalloreductase STEAP2 can be found in NCBI Protein database with accession numbers XP_043901516.1, XP_043891069.1, and XP_043891055.1, respectively.

Predicted mRNA sequences for *Solea senegalensis* pleiomorphic adenoma gene 1 (plag1), v-mos Moloney murine sarcoma viral oncogene homolog (mos), and STEAP family member 2, metalloreductase (steap2), can be found in NCBI Nucleotide database with accession numbers XM_044045581.1, XM_044035134.1, and XM_044035120.1, respectively.

Details of the organisms homologous to the selected protein can be found on NCBI Protein database with the provided accession numbers: *Micropterus salmoides* (XP_038579507.1), *Morone saxatilis* (XP_035509014.1), *Gadus morhua* (XP_030204512.1), *Thunnus albacares* (XP_044196375.1), *Mugil cephalus* (XP_047424079.1), *Oncorhynchus mykiss* (XP_021476238.1), *Xiphias gladius* (XP_039982848.1), *Chelmon rostratus* (XP_041817593.1), *Echeneis naucrates* (XP_029385537.1), *Anabas testudineus* (XP_026219050.1), *Spheniscus magellanicus* (KAF1422982.1), *Perca fluviatilis* (KAF1372749.1), *Atractosteus spatula* (MBN3317059.1), *Bos taurus* (NP_001071315.1), *Papio Anubis* (NP_001162460.1).

Paralogs within *Solea senegalensis* can be found in NCBI Protein database with the provided accession numbers: metalloreductase STEAP3-like (XP_043874980.1), metalloreductase STEAP3 (KAG7503596.1), metalloreductase STEAP4 (XP_043890536.1), metalloreductase STEAP4-like (XP_043871487.1), metalloreductase STEAP4 (KAG7475364.1), TNFA-induced adipose-related protein (ACO58567.1)

SUPPLEMENTARY DATA

Supplementary data is included with the submission. For additional information, please reach out to the corresponding authors.

ACKNOWLEDGEMENT

We would like to acknowledge Jonathan Bennett for his time and effort to respond to our questions.

FUNDING

No funding to report.

CONFLICT OF INTEREST

No conflicts of interest to report.

REFERENCES

1. *Solea senegalensis summary page*. FishBase. (n.d.). Retrieved October 16, 2022, from https://www.fishbase.se/summary/Solea-senegalensis.html
2. Centers for Disease Control and Prevention. (2020, September 16). *CDC - Anisakiasis - Frequently Asked Questions (FAQs)*. CDC. Retrieved October 16, 2022, from https://www.cdc.gov/parasites/anisakiasis/faqs.html
3. *What kind of organisms do Solea senegalensis host*. Global Biotic Interactions. (n.d.). Retrieved October 16, 2022, from https://www.globalbioticinteractions.org/?interactionType=hostOf&sourceTaxon=Solea+senegalensis
4. Guerrero-Cózar, I., Gomez-Garrido, J., Berbel, C., Martinez-Blanch, J. F., Alioto, T., Claros, M. G., Gagnaire, P. A., & Manchado, M. (2021). Chromosome anchoring in Senegalese sole (*Solea senegalensis*) reveals sex-associated markers and genome rearrangements in flatfish. Scientific reports, 11(1), 13460.
5. Ohgami RS, Campagna DR, McDonald A, Fleming MD. (2006). The Steap proteins are metalloreductases. Blood. 108(4):1388–1394. doi:10.1182/blood-2006-02-003681
6. Metalloreductase Family - Creative Biolabs. wwwcreative-biolabscom. [accessed 2022 Oct 14]. https://www.creative-biolabs.com/metalloreductase-family.html
7. U.S. National Library of Medicine. (n.d.). *Metalloreductase steap2 [Solea senegalensis] - protein - NCBI*. National Center for Biotechnology Information. Retrieved October 16, 2022, from https://www.ncbi.nlm.nih.gov/protein/XP_043891055.1
8. Saikia S, Oliveira D, Hu G, Kronstad J. (2013) Role of Ferric Reductases in Iron Acquisition and Virulence in the Fungal Pathogen Cryptococcus neoformans. Deepe GS, editor. Infection and Immunity. 82(2):839–850. doi:10.1128/iai.01357-13.
9. Yang Q, Ji G, Li J. (2019) STEAP2 is down-regulated in breast cancer tissue and suppresses PI3K/AKT signaling and breast cancer cell invasion in vitro and in vivo. Cancer Biology & Therapy. 21(3):278–291. doi:10.1080/15384047.2019.1685290
10. Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, proteomics & bioinformatics*, *2*(4), 216–221. https://doi.org/10.1016/s1672-0229(04)02028-5

11. Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *JMB, 268*(1), 78–94. https://doi.org/10.1006/jmbi.1997.0951
12. Clustal Colour Scheme. (n.d.). Retrieved October 14, 2022, from https://www.jalview.org/old/v2_8/help/html/colourSchemes/clustal.html
13. Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75
14. Waterhouse A.M., Procter J.B., Martin D.M.A., Clamp M., Barton G.J. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191. Pubmed: 19151095 DOI: doi:10.1093/bioinformatics/btp033
15. *About*. Jalview. (n.d.). Retrieved October 14, 2022, from https://www.jalview.org/about/

TABLE AND FIGURES LEGENDS

Table 1. Gene Prediction Using FGNESH and GENSCAN
Three gene prediction tools used for the analysis. Alignments are color-coded according to Blastp comparison results.

Figure 1. Multiple Sequence Alignment Using Clustal Omega
Multiple sequence alignment of *Solea senegalensis* metalloreductase STEAP2 and fifteen homologs. Alignment calculated using Clustal Omega. MSA display generated by Jalview using Clustal X color scheme.

Figure 2. UPGMA Phylogeny Tree
The UPGMA tree generated in MEGA11 shows the relationship between *Solea senegalensis* Metalloreductase STEAP2 and homologs in different species. A total of 16 sequences were analyzed and evolutionary pairwise distances were computed using the Poisson method. Pairwise deletion was implemented to account for gaps and missing data, resulting in a total of 535 positions in the dataset.

SUPPLEMENTARY TABLE AND FIGURES LEGENDS

Supplementary Figure 1. FGNESH Output
Gene prediction output from FGNESH 2.6 web version using *Austrofundulus limnaeus* parameter.

Supplementary Figure 2. GENSCAN Output
Gene prediction output from GENSCAN 1.0 web version using the vertebrate parameter.

Supplementary Figure 3. Pairwise Distances
Calculated Pairwise Distances using the Poisson method, demonstrating evolutionary divergence between 16 sequences.