Author: Christopher Tran

Goal: Examined MC1R gene by pulling data from the ensembl website using API with the given scripts.

Imported tools: requests, sys, re, from bio import SeqIO

References/database used: http://rest.ensembl.org/?content-type=text/html

Files created: MC1R.fasta, MC1R.py, MC1R\_homology\_list.txt, readme.txt, MC1R\_sequence.fasta

Execution: python3 MC1R.py

Instructions: MC1R.py was created to run the RESTful API and acquire data for MC1R. Below are step by step procedures to walk you through what is going on within the script.

Scripts were adjusted to have the script run successfully as most of the script was taken from the ensembl database.

1. Variables were set in ensure an easier and smoother script during the coding process.

Other variables were set along the script at the earliest point due to data not being generated at that stage. During the script, print statements placed to inform the audience the steps the script is running.

Using the mygene and ensembl database, the necessary API scripts were used throughout the script to find the ensembl id, DNA sequences, and homologous species

2.The first objective is to find the ensemble ID.

Using mygene data base, MC1R was searched within the database by the scripts that calls for a request to gather information from the database within loops.

Variable was set along with a json message function.

A print function was added to show the output of the data.

Determining the output was a dictionary and using [‘string’] id was able to be located and isolated.

The first id was 4157 which was not the ensembl id. This id made it one step closer to finding the ensemble id

3. Using the mygene server with a different api to locate the ensembl ,the id 4157 was used to generate another search.

The code to find the ensemble id was very similar to finding id 4157.( print and locating string from dictionary)

This code was also adjected to ensure the correct information was pulled.

The output generated ensembl id ENSG00000258839.4. A variable was set once the ensembl gene was found

4. The sequence of the gene was obtained next through the ensembl website through an API script.

Still using the request command the sequence was obtained.

Two fasta files were created and the sequence was written into both files

The files were named MC1R.fasta and MC1R\_sequence.fasta

MC1R.fasta will be used as the original file for this assignment.

MC1R\_sequence.fasta will be used strictly for the script run as after appending the the amino acid sequence the fasta file can not be run again.

5. The next objective was to find the longest open reading frame or ORF.

To do so a script was used to locate the ORF by using from bio import seqio or biopython.

With for loops, if and else statements, len command(count of the longest orf) and length restriction(which in this script the ORF had to be greater than 1300. This number can change depending on the amount you want), and regex to find the start codons of the ORF.

The ORF found was in a DNA sequence. The sequence was appended to the MC1R.fasta file with headers.

6. Once that was located, the sequence was converted to an amino acid.

The script for this is similar to the DNA sequence, but a amino acid table in biopython was used to convert the codons to the amino acid sequence.

Regex was not used in this portion of the script, but len, for loops, if statements and a length restriction(1300 which is similar to the DNA sequence) was used.

This amino acid sequence was also appended to the MC1R.fasta file with headers.

7. The last objective was to find a list of homologous species.

A script was located and used from the ensembl website.

Similar to the other ensembl scripts, a requested was used to go into the ensembl data base.

The API was used, and a list of species was also generated to see what was homologous to MC1R.

From the list, Regex was used to isolate the species and a loop was created to create and append MC1R\_homolgy\_list.txt to have the list of species within the file.

Below, you will see the results of the script

Results:

>ENSG00000258839.4 chromosome:GRCh38:16:89912119:89920973:1

TGAGGGCAGTGCCCAGAATTGAAGGCGAAGCCCCAGAAGCATGTTTTGCAGAGAAGTGCC

CAGGGAAGCTCTGAGGGCCCATGTAGCAAAGATCAGGGGATAGTCGGTCTGAGGGTGAAT

GGGCCACTCGGACCAAGACCCCAGTCTTGGGGGAGGGCTTAGCTGGAGCAGGTCCTGGCA

CAGTTGACTGATGGTGCACAGAACCCGTGCATCCCACGGCCCCACGGTGCTGCAGCTGCA

GGAGGGGCGGAGGCTGCAGCCAGACAGCATCAGAAGCCAGCGTGGTTCTGGAAGGATCGA

GAACACCAAGGTGTTAGGGCTGCAGCAGGGGTCCTGTCCCCTGGCACCCCTCACCGCCCT

AATCTTTTACCCTTAGGAGGCAGCAGACACGAGGGGCTGCCCGAGGCTCTAGGGCGGCCA

GTGAGGCAGGAAACATGTTCCAGCCCCAGCTAGGTACTGGTCCGTGGACCCACCTCCCAG

AAAGCCCATCACTGTGTAATCGTCTAACCTGGGGCTCGCCGAGGCCTGTGAGTTCATCCT

TTTGGCAGTTCCTGGTGTCTCCTTACTCTGCTCAGCATTTCCTGGGCGGGAGCTTAGGGT

GCAGGACCCTCCCCAGGACGACGAGGGCCCAGTGTCCATGACAAGAGTTGGCCCGAGGGC

TGAGCCACGTGTGCCCATCTCAGACGTGGGCCTGAGGGTGCAGCCCTGGCCCTGTGCTGG

CCATTTCTAGGAGCGGTGCCCTGAGGTCCCAGCTGTGATAGCCCCACGCTCTGCAGGAAG

AGATCATGGGGGCGGGGAGTTGGTGCTGCGGCCTCGTTCCTCTCTGCAGTGAGTGAACGA

TGTTTGTGGTCAGCAGGAGCCTGTGGGGAGCACAGGCTGGTCCTCCTGGTGTCCCACCCA

CCCCTTTTTCCATGGGGGATCTGCACTCATCTCCAGGGAAGATGGTTGGGAGATAACCCC

AGTCTGCTCTAGGTCCCCACCCTCCACAGCCAGGGTGGTCCGTGGTGAGCTTCAGCCATC

GAGATGCGGGAGTCTGCTAGAGTCTTCAGGGTCTTTTCTCTGAAAATGACAGGCTAGCAA

GGAGACCTGGGTCCCCTGCCTCTTCCATTCCAGATGCCTTGAGTCCACCCAAATAGGGGA

TGTGATGTTTGGAGCTGCAGCAGCCGCCCTACGGTTGGGAGTCAGAGAAGAGCCGGTGTT

CCAGGGACAATGCAGCAGAGGCTGAGCCCAGGCCTGCTGTCCTGAGAGGTGGCTGGATCA

CTGACACTTTGGCAGTGGTGCTGGGGTTTATGTCATGACCTGCAGCTGAGCCTACTTCCA

ATGACCGTGAGATCTGAAAGACTGTTTTGAGGGCGTAGCCTCTGCCATGATTGTGGGGAA

TGCTGTCCTGTTTCCTCCCTTGGCCCTGCTCAGCCCAGCGAGAGGCTGAGGCGCACGTGG

CTCCCCGGGTGCCCACAGGCAGCGTGGCTCACCAGCCGGGCCCTTTTCCACTGAGCCAGA

ACCCCCCAAAGCCTTCAATGCAGGCACCACGGTGAGCCCACGAGAAACCCTGCTTGCCAC

CTCCCACACCCCCACCCCCAAGTTCAAAGGAAATGGTCCCTGAACCAAGGGCTGAGATCA

GCTGTGGGTCCAGCTGTCCTGGGGAGCTGTACTGGAGCCCACCACGGTGGGACTGTTGGT

CCGGCGGTGACCCCCACCTCCATGTCTGTGGCCGCAGCTGGACAGGCCACTCCCTGGGCC

ACAGAGATGTTTTACCTCTCGCAGCCCTCGGGCACACATTGAGCAGATGTGTGTGTGTGT

GCGTGTGTGGGTGGGTGCGCATTTGTGTGTGCCTGTGTGTGTGCGCATGTGGTGTGGGTG

CACGTGTGTGCACGTGTGTGGGTAAACATTGTGTGTGCGCATACGTGTGTGGGTAAACAT

TTGTATGTGCACGCGTGTGTGGGTAAACGTGTGTGCGCACGTATGTATGTGTGTACATTT

GTATGTGTGTACATTTGTATGCATGTGTGCCTGTGTGTGTGTGCCTGTGTGTGTGGGTGC

ACATTTGTGTGTGTGTGTGTGCCTGTGTGTGTGTGTGCACGTATGTATGTGTGCACACTT

GTATGCATGTGTGCCTGTGTGTGTGGGTGCACATTTGTGTGTGTGTGCCTGTGTGTGTGG

GTGCACATTTGTGTGTGTGGGTGCACATTTGTGTGTGTGTGCGCCTGTGTGGGTGCACAT

TTGTGTGTGTGTGTGCCTGTGTGTGTGCCTGTGTGTGTGGGTGCCTGTGTGTGTGGGGCA

CATTTGTGTGTGTGTGTGCCTGTGTGTGGGTGCACATTTGTGTGTGTGCCTCTGTGTGTG

TGCCTGTGTGTGGGGGTGCACATTTGTGTGTGCGCCTGTGTGTGGGGGTGCACATTTGTG

TGTGCGCCTGTGTGTGTGGGTGCACATTTGTGTGTGTGTGCGCCTGTGTGTGTGGGTGCC

TGTGTGTGTGTGGGGCACATTTGTGTGTGTGTGTGTGCCTGTGTGTGGGTGCACATTTGT

GTGTGTGCCTGTGTGTGTGTGCCTGTGTGTGGGGGTGCACATTTGTGTGTGTGTGTGCCT

GTGTGTGGGGGTGCACATTTGTGTGTGTGTGTGCCTGTGTGTGTGGGTGCACATTTGTGT

GTGTGTGTGCCTGTGTGTGTGGGTGCACATTTGTGTGTGTGTGCCTGTGTGTGTGGGTGC

ACATTTGTGTGTGTGTGTGCCTGTGTGTGTGGGTGCACATTTGTGTGTGTGCCTGTGTGT

GGGTGCACATTTGTGTGTGTGTGCCTGTGTGTGTGTGCCTGTGTGTGTGGGTGCACATTT

GTGTGTGTGTGTGCCTGTGTGTGTGGGTGCACATTTGTGTGTGTGTGTGTGCCTGTGTGT

GTGTTGCAGGCCCTGGATGCCAGACACTGAATAAACGCAGGAAGACGTCTGTCTTCATTC

TCCTCGTGGGTCGCTGGTCCAGAAACACCTGGATGGAAAGTGCTCTGCAGGAACGGTGCC

TCTGCCTGTGGCGGGGACCCTGGTGAGCGGATGGGCCAGCCCCACGTGTCTTCCGGCCAC

TCAGCATGCAGTGTTTCCAGGGGCACTAAGAGACCAAAATCGAGATATGATTAGCTGTAG

GATGTCATCTAATCACAGATCATCCCGAGGCTAATTTATCTCCCCCATGACCATAACACA

TCAAAAAGTTGACTTTTTGCAGCTCGGCTGTGCCTCATCTTCCCACGAAGCCCCGACAGG

CACATCCAGTGAGGAACCACAGTGGGAGTCCTGTGGCAGGGTCACCCCACTTCCGATGCC

CTCCAGCTGCATCTTGGCACGAAAAAGGCTGCCCAGTTCTCATGCCCTTTCAAGTCCCGG

GCTGGGGGGTAGCAGTGGAGGCTGGGGTTAACCGCCGTCCGTCTCAAAGGCCTCATTGTG

GAGCTGCAAACACGAACGTCCTTGAAATGTGAGGGGACAGTGCTCTGGGGCAGGGGCTGC

TCACTGGTTCATGGCCAGAGGTGAGCGGGCTCTGGGTCTGGGTCTGTGAGGTGCTGTGGA

CGTCGGGGGTGCTCCTGGGGCAGGGACACAGCCACGGCCCTCACACCAGTGGAGCCGTCT

TCCTTCCCGAGGCAGAGGCTTGGCCTTCTCACACCTTGGGACCCTTCTCTCCCCTGTGCT

CATGCTGGAAGCACAGCGTGAGGGGGCAAAGGTCATGGGAGGCAAAAGGCTGGGCTGAGG

CCGAGGCTCTGTGGCTGTGGCCGGATACCAGGTCCTGTGGTGGTGTGGGCAGGGCGCGCT

CTCTCCTCTGAGCAGCCTGGGGCTGCGTGTGTGAACAGAAACAGGCCTGCCGATTTGAAG

ATGGTTTGACCAAGTCTCCCAAAGCTGAAGAGAGGTCCCTGCCCATGTCCCCCGGGTGGC

AATCCACAGACGCACCTGTGTCCATCAGACGGTATCGTTGATAGGGAGTCTGAAGCCACC

CAGACGTGGTCTGTTCACGCCGTGGACACCACGGTCTGAGGCAGACGGTTACAGCCACCG

GCATAGGCGAATCTCAGCCAGAGAGAGTGAACTGGGACACCATTCACGTGATGGGCCAGA

TGGGCTGGCAGCGGCTCGCTGGAGCCGTCCGGGTGGGAACGCTGTGTCTGTTGGTCTGGT

GCTGCTTATGTGGCTGGTTCAGGTCTGTCATCCGTCAACCTGCATATTTATCACTGGTGC

ATTTTCATGTATGTTGTACCTCAATTAAAACATTTTAGGGCCGGGCGCGGTGGCTCACGC

CTGTAATCCCAGCACTGTGGGAGGCGGAGGCGGGAGGATCATGAGGTCAGGAGATCAAGA

CCGTCCTGGCCAACATGGTGAAACCCCATCTCTACTAAAAAAAATACAAAAATTTGCCGG

GTGTGGCGGCACGCGCCTGTAGTTCCAGCTACTCAGGAGGCTGAGGTATGAGAATTACTT

GAACCCAGTAGGCAGAGGCTACCGTGAGCCAAGATCATGCCACTGCACTCCAGCCTGGGT

GACAGTGAGACTTCGTCTCAAAAAAAAAAAAAAAAAAATTAAAGCAGATGGGGTGTGGGG

GCTCATCCCTGTAATCTCAGCATTTTGGGAGGCTGAGGCGGACCGATCACCTGAGGTCAG

GAGTTCAAGACCAGCCTGGTCAACATGGTGAAACCTCATCTCTACAAAAAATACAAAAAA

TAGCCAGGTGTGGTAGTGGGTGCCGTAGTCCCAGCTAGTTGGGAGGCTGAGGCATAAGTA

TTGCTTGAATCCAGAAGGTGGAGGTTGCAGTGAGCCAAGAATGCACCACTACACTCCAGC

CTGCGCGACAGAGTGAAACTGTCTCAAAATAAATAAATACATAAAAATTATATGTATATA

TATATATATATATATTTTTTTTTTTTTGAGACTGTATCTCTGTTTCCCAGGCTGGAGTGC

AGTGGTGTGATCTCGGCTCACTGCAACCTCCGCCTCCTGGGTTCAAGCAATTCTGCCTCA

GCCTCCCAAGTAGCTGGGATTATAGGCGTGCGCCACCACGCCTGGCTGATTTTTGTATTT

TTAGTAGAGATGGAGTTTCACCATATTACCCAAGCTGGTCTTGAACTCCTGACCTCATGA

TCCGCCTGCCTCAGCCTCCCAAAGTGCTGAGATTATAGGGGTGAGCCACTGTGCCCGGCA

ATAATAATTTTTTTTTTTTTTGAGACAGAGTTTTGCTCTTGTTGCCCAGGCTGGAGTGCA

ATGGTGCAATCTCGGCTCACGGCAACCTCCGCCTCTCATGTTCAAGCTATTATCCTGACT

CAGCTGGGATTACAGGCACCCACCACCACACCTGGCTGATTTTGGTATTTTTAGTAGAGA

CGGGGTTTCTCCATGTCGGCCAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCCACCCG

CCTTGGGCTCCCAAAGTGCTGGGATTACAGCCCTGAGCCACCATGGCTGACAATAATAAA

AAATTTTAAAGCAAAACTAAAGATTTATAAATTGTAAAATGCGGTTGTAACAAGTATTTA

CTGTACCAGTAAATATGAAAAATATTAATGTCCTCACACACATGGAGGTGGCTTGTGAGT

GGTGTGAGCATTTCATGGTCAGAGTTTGGCAGTTGGGCGAGAGACAGGATGGATTTTAAC

TCGAACATCGCAGACACAAAACTGAAAACCCTGATGCGGTGCCTGCACCGTTGCCTCCTT

TCCTGCAGCTGTAAATGGGACGAGTAATTCCATCCCCCTCCCGCTTCCACCCTTCAGCAC

AGACGCAGTCTTCAGCAAGGAAGTGCTGGGAACGCCCTGGAGTGAACCCAGGAAGATGCC

TGCAGTGGGTGCCAGGGCCCCTCTCCACCGTCCCTGCTGGGCTTCGGGGCCACGCCCGAC

TGCTGTGAACGGCCTGCGGAGCACCACGTGCGACGGCTGGAGGCGAGAGGTCTGCCTTTG

ATGTGGCTGTTGGTGCAGGGCCTGTGGTGCCTTCCGCAGCGGAAATGGCGCGCCGCCCGG

GGAGGGCGGGAGCAGCGTCCCGGGTGCCCCTGTGAGGATGAGCGACGAGATGACTGGAGG

GTCCCTGAAGACCTCACTAGGGTGCCCCCAGCCGGTCCGCTCCCAGGAAGCGACACCCCC

ACAGCCCCAGGGCTGCAGCTGAGGGGGTCGCCACTCTGGCTGGGCGAGGCTGGGCCCTTG

GGGGCAGGCGCCAGAGTGGCCTCAGGCTCTACAAGATGCCTGAAAACACCAACCTCTCCA

GGGCTCACTAGCATTGGACGCTTTCACGCTCTGCCCTGGCCGGAAGCCCCCTCACCCCGC

GCGATGTGCAAACTCCTGCAGGGCTCACTCAGTTTCCAGAACTTTAATTATTGGAAAGTT

CTCCCTGGTCCAGCCCCCAAATCTGCCGTGAACGTTGACAGCTGAGTTGCTGCTCCATGC

GTGCTTTGGCTGAGAGCAGAGGGGACCCCTGTCCTCCCTGAGCTGCTGACGAGGGGAGGG

GTGAAGGGTGGGGCCTCTGGAGAGGGCAGGTCCCGGGGAAGCTCCGGACTCCTAGAGGGG

CGGCCAGGTGGGGGCCCTGGTGACCAGGACAGACTGTGGTGTTTTTTAACGTAAAGGAGA

TCCGCGGTGTGAGGGACCCCCTGGGTCCTGCACGCCGCCTGGTGGCAGGCCGGGCCATGG

TGGGTGCTCACGCCCCCGGCATGTGGCCGCCCTCAGTGGGAGGGGCTCTGAGAACGACTT

TTTAAAACGCAGAGAAAAGCTCCATTCTTCCCAGGACCTCAGCGCAGCCCTGGCCCAGGA

AGGCAGGAGACAGAGGCCAGGACGGTCCAGAGGTGTCGAAATGTCCTGGGGACCTGAGCA

GCAGCCACCAGGGAAGAGGCAGGGAGGGAGCTGAGGACCAGGCTTGGTTGTGAGAATCCC

TGAGCCCAGGCGGTAGATGCCAGGAGGTGTCTGGACTGGCTGGGCCATGCCTGGGCTGAC

CTGTCCAGCCAGGGAGAGGGTGTGAGGGCAGATCTGGGGGTGCCCAGATGGAAGGAGGCA

GGCATGGGGGACACCCAAGGCCCCCTGGCAGCACCATGAACTAAGCAGGACACCTGGAGG

GGAAGAACTGTGGGGACCTGGAGGCCTCCAACGACTCCTTCCTGCTTCCTGGACAGGACT

ATGGCTGTGCAGGGATCCCAGAGAAGACTTCTGGGCTCCCTCAACTCCACCCCCACAGCC

ATCCCCCAGCTGGGGCTGGCTGCCAACCAGACAGGAGCCCGGTGCCTGGAGGTGTCCATC

TCTGACGGGCTCTTCCTCAGCCTGGGGCTGGTGAGCTTGGTGGAGAACGCGCTGGTGGTG

GCCACCATCGCCAAGAACCGGAACCTGCACTCACCCATGTACTGCTTCATCTGCTGCCTG

GCCTTGTCGGACCTGCTGGTGAGCGGGAGCAACGTGCTGGAGACGGCCGTCATCCTCCTG

CTGGAGGCCGGTGCACTGGTGGCCCGGGCTGCGGTGCTGCAGCAGCTGGACAATGTCATT

GACGTGATCACCTGCAGCTCCATGCTGTCCAGCCTCTGCTTCCTGGGCGCCATCGCCGTG

GACCGCTACATCTCCATCTTCTACGCACTGCGCTACCACAGCATCGTGACCCTGCCGCGG

GCGCGGCGAGCCGTTGCGGCCATCTGGGTGGCCAGTGTCGTCTTCAGCACGCTCTTCATC

GCCTACTACGACCACGTGGCCGTCCTGCTGTGCCTCGTGGTCTTCTTCCTGGCTATGCTG

GTGCTCATGGCCGTGCTGTACGTCCACATGCTGGCCCGGGCCTGCCAGCACGCCCAGGGC

ATCGCCCGGCTCCACAAGAGGCAGCGCCCGGTCCACCAGGGCTTTGGCCTTAAAGGCGCT

GTCACCCTCACCATCCTGCTGGGCATTTTCTTCCTCTGCTGGGGCCCCTTCTTCCTGCAT

CTCACACTCATCGTCCTCTGCCCCGAGCACCCCACGTGCGGCTGCATCTTCAAGAACTTC

AACCTCTTTCTCGCCCTCATCATCTGCAATGCCATCATCGACCCCCTCATCTACGCCTTC

CACAGCCAGGAGCTCCGCAGGACGCTCAAGGAGGTGCTGACATGCTCCTGGTGAGCGCGG

TGCACGCGGCTTTAAGTGTGCTGGGCAGAGGGAGGTGGTGATATTGTGTGGTCTGGTTCC

TGTGTGACCCTGGGCAGTTCCTTACCTCCCTGGTCCCCGTTTGTCAAAGAGGATGGACTA

AATGATCTCTGAAAGTGTTGAAGCGCGGACCCTTCTGGGTCCAGGGAGGGGTCCCTGCAA

AACTCCAGGCAGGACTTCTCACCAGCAGTCGTGGGGAACGGAGGAGGACATGGGGAGGTT

GTGGGGCCTCAGGCTCCGGGCACCAGGGGCCAACCTCAGGCTCCTAAAGAGACATTTTCC

GCCCACTCCTGGGACACTCCGTCTGCTCCAATGACTGAGCAGCATCCACCCCACCCCATC

TTTGCTGCCAGCTCTCAGGACCGTGCCCTCGTCAGCTGGGATGTGAAGTCTCTGGGTGGA

AGTGTGTGCCAAGAGCTACTCCCACAGCAGCCCCAGGAGAAGGGGCTTTGTGACCAGAAA

GCTTCATCCACAGCCTTGCAGCGGCTCCTGCAAAAGGAGGTGAAATCCCTGCCTCAGGCC

AAGGGACCAGGTTTGCAGGAGCCCCCCTAGTGGTATGGGGCTGAGCCCTCCTGAGGGCCG

GTTCTAAGGCTCAGACTGGGCACTGGGGCCTCAGCCTGCTTTCCTGCAGCAGTCGCCCAA

GCAGACAGCCCTGGCAAATGCCTGACTCAGTGACCAGTGCCTGTGAGCATGGGGCCAGGA

AAGTCTGGTAATAAATGTGACTCAGCATCACCCAC

>ENSG00000258839.4's Longest ORF DNA sequence: ATGCTG...TGA - length(nucleotide): 1599, strand: -1, frame: 1, pos: 0

ATGCTGAGTGGCCGGAAGACACGTGGGGCTGGCCCATCCGCTCACCAGGGTCCCCGCCACAGGCAGAGGCACCGTTCCTGCAGAGCACTTTCCATCCAGGTGTTTCTGGACCAGCGACCCACGAGGAGAATGAAGACAGACGTCTTCCTGCGTTTATTCAGTGTCTGGCATCCAGGGCCTGCAACACACACACAGGCACACACACACACACAAATGTGCACCCACACACACAGGCACACACACACACAAATGTGCACCCACACACACAGGCACACACACACAGGCACACACACACAAATGTGCACCCACACACAGGCACACACACAAATGTGCACCCACACACACAGGCACACACACACACAAATGTGCACCCACACACACAGGCACACACACACAAATGTGCACCCACACACACAGGCACACACACACACAAATGTGCACCCACACACACAGGCACACACACACACAAATGTGCACCCCCACACACAGGCACACACACACACAAATGTGCACCCCCACACACAGGCACACACACACAGGCACACACACAAATGTGCACCCACACACAGGCACACACACACACACAAATGTGCCCCACACACACACAGGCACCCACACACACAGGCGCACACACACACAAATGTGCACCCACACACACAGGCGCACACACAAATGTGCACCCCCACACACAGGCGCACACACAAATGTGCACCCCCACACACAGGCACACACACAGAGGCACACACACAAATGTGCACCCACACACAGGCACACACACACACAAATGTGCCCCACACACACAGGCACCCACACACACAGGCACACACACAGGCACACACACACACAAATGTGCACCCACACAGGCGCACACACACACAAATGTGCACCCACACACACAAATGTGCACCCACACACACAGGCACACACACACAAATGTGCACCCACACACACAGGCACACATGCATACAAGTGTGCACACATACATACGTGCACACACACACACAGGCACACACACACACACAAATGTGCACCCACACACACAGGCACACACACACAGGCACACATGCATACAAATGTACACACATACAAATGTACACACATACATACGTGCGCACACACGTTTACCCACACACGCGTGCACATACAAATGTTTACCCACACACGTATGCGCACACACAATGTTTACCCACACACGTGCACACACGTGCACCCACACCACATGCGCACACACACAGGCACACACAAATGCGCACCCACCCACACACGCACACACACACACATCTGCTCAATGTGTGCCCGAGGGCTGCGAGAGGTAAAACATCTCTGTGGCCCAGGGAGTGGCCTGTCCAGCTGCGGCCACAGACATGGAGGTGGGGGTCACCGCCGGACCAACAGTCCCACCGTGGTGGGCTCCAGTACAGCTCCCCAGGACAGCTGGACCCACAGCTGATCTCAGCCCTTGGTTCAGGGACCATTTCCTTTGAACTTGGGGGTGGGGGTGTGGGAGGTGGCAAGCAGGGTTTCTCGTGGGCTCACCGTGGTGCCTGCATTGAAGGCTTTGGGGGGTTCTGGCTCAGTGGAAAAGGGCCCGGCTGGTGA

>ENSG00000258839.4 Longest ORF Amino Acid: MLS...PGW - length(nucleotide): 1599, length(amino acid): 533.0, strand: -1, frame :1, pos: 5789

MLSGRKTRGAGPSAHQGPRHRQRHRSCRALSIQVFLDQRPTRRMKTDVFLRLFSVWHPGPATHTQAHTHTQMCTHTHRHTHTQMCTHTHRHTHTGTHTQMCTHTQAHTQMCTHTHRHTHTQMCTHTHRHTHTNVHPHTQAHTHTNVHPHTQAHTHTNVHPHTQAHTHTNVHPHTQAHTHRHTHKCAPTHRHTHTHKCAPHTHRHPHTQAHTHTNVHPHTQAHTQMCTPTHRRTHKCAPPHTGTHTEAHTQMCTHTQAHTHTNVPHTHRHPHTQAHTQAHTHTNVHPHRRTHTQMCTHTHKCAPTHTGTHTQMCTHTHRHTCIQVCTHTYVHTHTQAHTHTQMCTHTHRHTHTGTHAYKCTHIQMYTHTYVRTHVYPHTRAHTNVYPHTYAHTQCLPTHVHTRAPTPHAHTHRHTQMRTHPHTHTHTHLLNVCPRAARGKTSLWPREWPVQLRPQTWRWGSPPDQQSHRGGLQYSSPGQLDPQLISALGSGTISFELGGGGVGGGKQGFSWAHRGACIEGFGGFWLSGKGPGW

Below are the unique homologous species:

ornithorhynchus\_anatinus

sarcophilus\_harrisii

dasypus\_novemcinctus

callithrix\_jacchus

macaca\_fascicularis

macaca\_mulatta

macaca\_nemestrina

balaenoptera\_musculus

ovis\_aries\_rambouillet

felis\_catus

sus\_scrofa

ursus\_americanus

mesocricetus\_auratus

panthera\_leo

ailuropoda\_melanoleuca

monodon\_monoceros

mus\_musculus

vombatus\_ursinus

catagonus\_wagneri

bison\_bison\_bison

mus\_spicilegus

bos\_taurus

marmota\_marmota\_marmota

physeter\_catodon

monodelphis\_domestica

microcebus\_murinus

mus\_caroli

cricetulus\_griseus\_chok1gshd

vulpes\_vulpes

neovison\_vison

bos\_mutus

moschus\_moschiferus

bos\_grunniens

rhinolophus\_ferrumequinum

capra\_hircus

sciurus\_vulgaris

ictidomys\_tridecemlineatus

ictidomys\_tridecemlineatus

phocoena\_sinus

delphinapterus\_leucas

aotus\_nancymaae

mus\_pahari

heterocephalus\_glaber\_female

nannospalax\_galili

cavia\_porcellus

mus\_spretus

carlito\_syrichta

phascolarctos\_cinereus

bos\_indicus\_hybrid

oryctolagus\_cuniculus

saimiri\_boliviensis\_boliviensis

peromyscus\_maniculatus\_bairdii

cervus\_hanglu\_yarkandensis

chlorocebus\_sabaeus

rhinopithecus\_bieti

ciona\_intestinalis

petromyzon\_marinus

eptatretus\_burgeri

eptatretus\_burgeri

callorhinchus\_milii

lepisosteus\_oculatus

clupea\_harengus

danio\_rerio

cyprinus\_carpio

ictalurus\_punctatus

electrophorus\_electricus

esox\_lucius

oncorhynchus\_kisutch

oncorhynchus\_kisutch

oncorhynchus\_mykiss

oncorhynchus\_mykiss

salmo\_trutta

salmo\_trutta

salmo\_salar

salmo\_salar

gadus\_morhua

fundulus\_heteroclitus

poecilia\_reticulata

xiphophorus\_maculatus

oryzias\_latipes

cyclopterus\_lumpus

oreochromis\_niloticus

haplochromis\_burtoni

astatotilapia\_calliptera

sparus\_aurata

lates\_calcarifer

xenopus\_tropicalis

chrysemys\_picta\_bellii

sphenodon\_punctatus

notechis\_scutatus

laticauda\_laticaudata

pseudonaja\_textilis

anas\_platyrhynchos\_platyrhynchos

meleagris\_gallopavo

gallus\_gallus

geospiza\_fortis

oryzias\_melastigma

chelonoidis\_abingdonii

acanthochromis\_polyacanthus

pygocentrus\_nattereri

aquila\_chrysaetos\_chrysaetos

takifugu\_rubripes

anser\_brachyrhynchus

coturnix\_japonica

amphiprion\_percula

nothobranchius\_furzeri

seriola\_dumerili

taeniopygia\_guttata

ficedula\_albicollis

salvator\_merianae

hippocampus\_comes

maylandia\_zebra

amphilophus\_citrinellus

gopherus\_evgoodei

naja\_naja

neolamprologus\_brichardi

dicentrarchus\_labrax

myripristis\_murdjan

struthio\_camelus\_australis

poecilia\_formosa

betta\_splendens

paramormyrops\_kingsleyae

sinocyclocheilus\_grahami

podarcis\_muralis

cynoglossus\_semilaevis

gasterosteus\_aculeatus

strigops\_habroptila

erpetoichthys\_calabaricus

stegastes\_partitus

anabas\_testudineus

kryptolebias\_marmoratus

cottoperca\_gobio

oncorhynchus\_tshawytscha

oncorhynchus\_tshawytscha

pundamilia\_nyererei

hucho\_hucho

hucho\_hucho

seriola\_lalandi\_dorsalis

scophthalmus\_maximus

sander\_lucioperca

amphiprion\_ocellaris

labrus\_bergylta

tetraodon\_nigroviridis

oryzias\_javanicus

oryzias\_sinensis

cyprinodon\_variegatus

terrapene\_carolina\_triunguis

terrapene\_carolina\_triunguis

poecilia\_latipinna

larimichthys\_crocea

scleropages\_formosus

scleropages\_formosus

denticeps\_clupeoides

leptobrachium\_leishanense

mastacembelus\_armatus