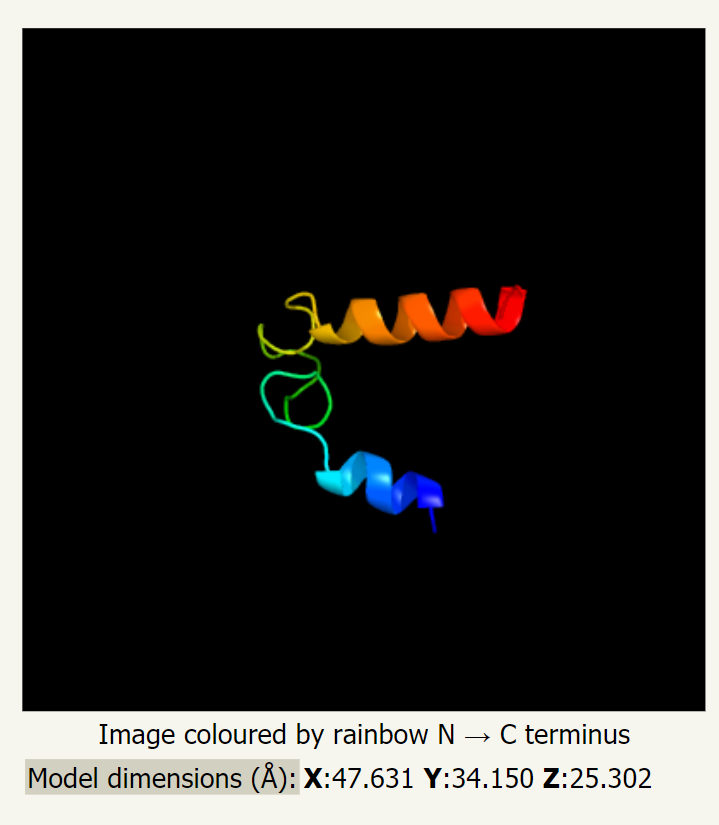
**Investigation of *ccdc190* gene using online tools and databases**



Guangxin Chen, GCD 3485

[chen6751@umn.edu](mailto:chen6751@umn.edu) 5509431

University of Minnesota

08/01/2021

Content

General information ------------------------------------------------ Page 3

Transcription regulation -------------------------------------------- Page 4 – 11

mRNA and translation regulation --------------------------------- Page 11 – 12

Protein structure and localization --------------------------------- Page 13 – 16

Post translation modification and protein interaction ---------- Page 17 – 22

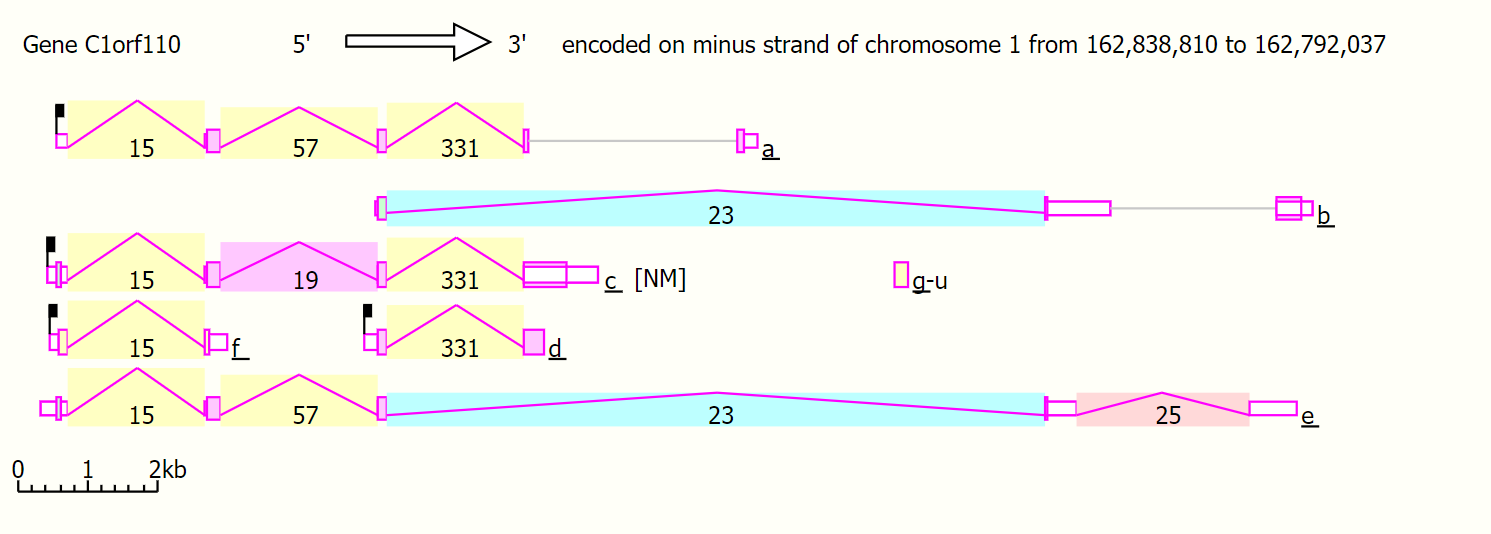
Homology and evolution ------------------------------------------- Page 23 – 28

Function and clinical significance --------------------------------- Page 29 - 30

General information

**Coiled-Coil Domain Containing 190,** also known as **C1orf110**, the Chromosome 1 Open Reading Frame 110, MGC48998 and CCDC190**,** is found to be a protein coding gene widely expressed in vertebrates. 1 Though not well-understood, it is found to be a protein coding gene. In human genome, it locates in 1q23.3 with a size of 44,461 bps. 2,3 The closest neighbors of ccdc190 gene are HSD17B7 and DDR2 gene in the long arm for chromosome 1. 4

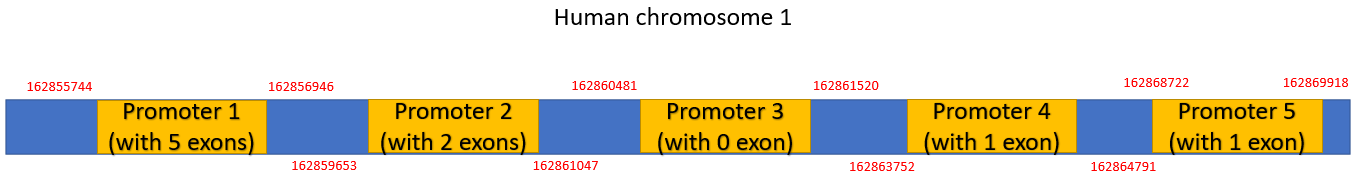
RNA-seq expression profile shows that this gene expressed most at lung in human body, followed by heart, brain, ovary and testis. 2 The expression product of *c1orf110* is often called Coiled-coil domain-containing protein 190 with a size of 302 aa.3 It may get the name because a coiled-coil domain is found from position 14 to 72. Two isoforms of such protein were identified which is caused by alternative splicing, while based on AceView, this gene can have at least 4 spliced variants of mRNA (Figure 1). 4



**Figure 1:** The alternative mRNAs aligned from 5' to 3', the region covered with different colors are the introns from AceView database. Exons are located at both ends of the intron. Their lengths can be estimated from the calibrator shown at bottom left. [[1]](#footnote-1)

Transcription regulation

The promoter regions of *ccdc190* gene are predicted based on the result generated from Genomatix (Figure 2). Based on the prediction, this gene has five promoters from 162855744 to 162869918 on chromosome 1.



**Figure 2:** Promoter regions of human *ccdc190* gene. Red numbers show the location of the start and ending base of the promoter in chromosome 1. Yellow areas represent the promoters with exons.

Also, the predicted transcription factor of promoters is labelled on the promoter sequence shown below. The chosen sequence is from predicted promoter 5 shown on figure 2. The transcription factors are found based on the output from Genomatix, factors with high matrix similarity are chosen and labelled (Table 1).

CCTTCCCTGATTAGCTTAATAACTAACCTCCTGAATTTTTTTTTCAGGTA 50  
GGAAGGGACTAATCGAATTATTGATTGGAGGACTTAAAAAAAAAGTCCAT  
  
  
AATCAGGGATTTCTTCTTGGTTT**GGA**TGCATTGCTGGTGAACTAGTGTGA 100  
TTAGTCCCTAAAGAAGAACCAAA**CCT**ACGTAACGACCACTTGATCACACT

Nuclear factor kappa B/c-rel

TTTGGGGGGAGTGCTGAAGAGCCTTGT**TTTGTC**ATATTACCAGGGTCGGT 150  
AAACCCCCCTCACGACTTCTCGGAACA**AAACAG**TATAATGGTCCCAGCCA  
  
  
TTTCTGGTTCCTTCTCATTTGGGGAGTCTC**TGTCAG**AGGGAAGGTCTAGG 200  
AAAGACCAAGGAAGAGTAAACCCCTCAGAG**ACAGTC**TCCCTTCCAGATCC  
  
  
GT**TGAAG**CCTGTTGTTCAGATTATTTTGTCTCACGGG**GTG**TTCCCTTGAT 250  
CA**ACTTC**GGACAACAAGTCTAATAAAACAGAGTGCCC**CAC**AAGGGAACTA  
  
  
GTAGTACTCTCCCCCTTTTCCTATGGATGTAACTTCCTGT**G**GGC**CAAA**CT 300  
CATCATGAGAGGGGGAAAAGGATACCTACATTGAAGGACA**C**CCG**GTTT**GA

Nuclear receptor subfamily 2 factors

Human and murine ETS1 factors

TALE homeodomain class recognizing TG motifs

Mouse Krueppel like factor

Nuclear factor kappa B/c-rel

Bicoid-like homeodomain transcription factors

Hepatic Nuclear Factor 1

GCAGTGATTGTTGTCTC**TCTT**CTGGGTCTAGCCACCCAGAGAGTCTACCT 350  
CGTCACTAACAACAGAG**AGAA**GACCCAGATCGGTGGGTCTCTCAGATGGA  
  
  
GGCTCCAGGCTAACACTGGGGGTTGTCTACAGAGTGTCCTGTGATGTGAA 400  
CCGAGGTCCGATTGTGACCCCCAACAGATGTCTCACAGGACACTACACTT  
  
  
CTGTCTGTGGGTCTCTCAGCCATGAATACCAGT**GTCTG**TTCCAGTGGAGG 450  
GACAGACACCCAGAGAGTCGGTACTTATGGTCA**CAGAC**AAGGTCACCTCC  
  
  
TGGTGGGGTGCACTGGACCCAGTGAGGGTTCTTAGCTTTGGTTGTT**TAAT** 500  
ACCACCCCACGTGACCTGGGTCACTCCCAAGAATCGAAACCAACAA**ATTA**  
  
  
GTTCTATTTCTGTGCTGGTTGGCCTCTTGCCTGAGGTGGCGCCTTCCAGA 550  
CAAGATAAAGACACGACCAACCGGAGAACGGACTCCACCGCGGAAGGTCT  
  
  
AAGCATCAGTTGTGGTAGTACTGGGG**GGAA**CTGGCAGTGGGTGGGGCCCT 600  
TTCGTAGTCAACACCATCATGACCCC**CCTT**GACCGTCACCCACCCCGGGA

Krueppel like transcription factors

PAX-4/PAX-6 paired domain binding sites

Lim homeodomain factors

Human and murine ETS1 factors

Hepatic Nuclear Factor 1

AGAACTCC**CAAGGTTA**TAAGCCCTTTTTCTTCTGCTACCAGGGTAGGTAG 650  
TCTTGAGG**GTTCCAAT**ATTCGGGAAAAAGAAGACGATGGTCCCATCCATC  
  
  
GGAAGGACCATCAGGTGGGGGCGGGGCTAGGCGTGTCTGAGCTCAGACTC 700  
CCTTCCTGGTAGTCCACCCCCGCCCCGATCCGCACAGACTCGAGTCTGAG  
  
  
TGCTT**GGGTGG**GTCTCGCTTGGCTGCTGTGGGGGATGGGGGTGAGGTTCC 750  
ACGAA**CCCACC**CAGAGCGAACCGACGACACCCCCTACCCCCACTCCAAGG  
  
  
CTTGTCAATAGAGTTGTGTACCCAGGATTATGGCTGCCTCT**GCTGAGTCA** 800  
GAACAGTTATCTCAACACATGGGTCCTAATACCGACGGAGA**CGACTCAGT**  
  
  
CGCAGGTTGTCAGGGAAGTGGGGGAAACCCAGCAGTCACAGGCCTCACCC 850  
GCGTCCAACAGTCCCTTCACCCCCTTTGGGTCGTCAGTGTCCGGAGTGGG  
  
  
AGCTCCCACACAAAATGAAGGGCTGGTCTCCCTCCCACCATGCCCCCCCG 900  
TCGAGGGTGTGTTTTACTTCCCGACCAGAGGGAGGGTGGTACGGGGGGGC

Fork head domain factors

MAF and AP1 related factors

Krueppel like transcription factors

Vertebrate steroidogenic factor

C**AACAA**CCCCAAGTCT**GTTT**CCAGGTGGAGGCTGTAACAGGGTTGAAAAC 950  
G**TTGTT**GGGGTTCAGA**CAAA**GGTCCACCTCCGACATTGTCCCAACTTTTG  
  
  
TTGCCCCAGGCTACCGGCTGCTAAAGAAAAGGGCTTGGTTCGTCCCCCAC 1000  
AACGGGGTCCGATGGCCGACGATTTCTTTTCCCGAACCAAGCAGGGGGTG  
  
  
**C**TGTGGAGTC**TGCACACC**AGATTTGCGCCC**TCCCC**TGAGTTCTGACCAGG 1050  
GACACCTCAG**ACGTGTGG**TCTAAACGCGGG**AGGGG**ACTCAAGACTGGTCC  
  
  
A**GGCGTC**TCGCCCCATCCAAATTGTTAAAAAGTTCA**G**CTGGACATTTCCC 1100  
T**CCGCAG**AGCGGGGTAGGTTTAACAATTTTTCAAGTCGACCTGTAAAGGG  
  
  
TCTCCCCATGGAGTT**C**TA**CCCC**TG**C**TCCTCTGGTCACCCTCCTGATGGAT 1150  
AGAGGGGTACCTCAA**G**AT**GGGG**AC**G**AGGAGACCAGTGGGAGGACTACCTA  
  
  
CCCCATGGTGCCAGGCAGGAATGGCCTGCTAGGAGATGCAGTGAGCC 1197  
GGGGTACCACGGTCCGTCCTTACCGGACGATCCTCTACGTCACTCGG

Testis-specific bHLH-Zip transcription factors

**transcription start region of RNA sequence from NCBI**

CT element binding protein

C2H2 zinc finger transcription factors 2

Metal induced transcription factor

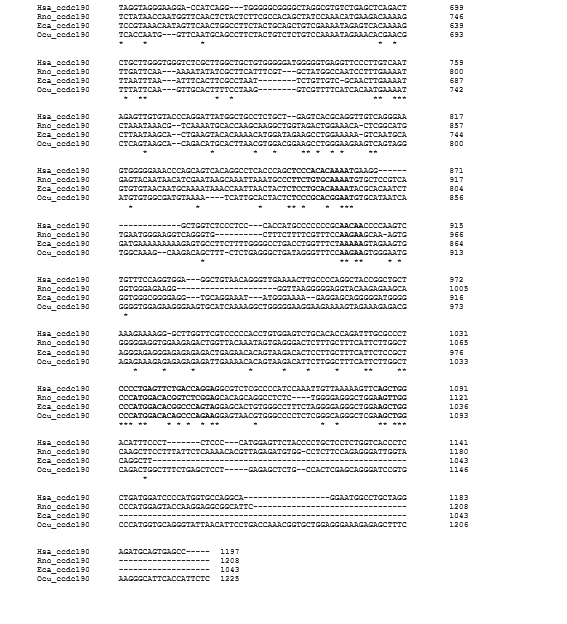
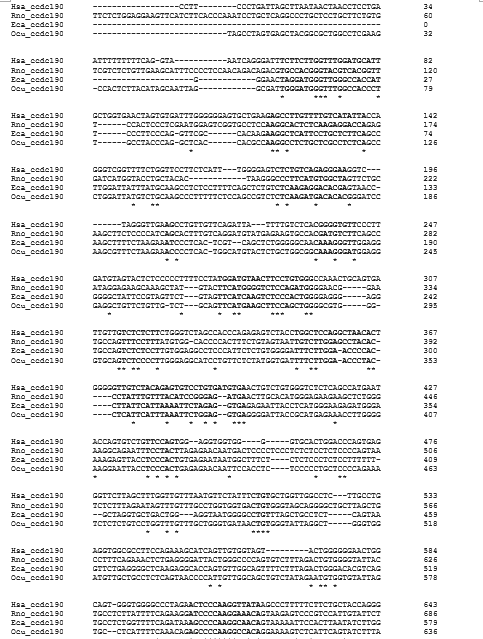
CP2-erythrocyte Factor related to drosophila Elf1

**transcri-ption start region of RNA sequence from Genomatix**

|  |  |  |
| --- | --- | --- |
| Matrix Family | Detailed Family information | Matrix sim. |
| [V$NFKB](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24NFKB) | Nuclear factor kappa B/c-rel | 0.916 |
| [V$MOKF](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24MOKF) | Mouse Krueppel like factor | 0.991 |
| [V$TALE](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24TALE) | TALE homeodomain class recognizing TG motifs | 1 |
| [V$BCDF](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24BCDF) | Bicoid-like homeodomain transcription factors | 0.948 |
| [V$NFKB](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24NFKB) | Nuclear factor kappa B/c-rel | 0.943 |
| [V$NR2F](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24NR2F) | Nuclear receptor subfamily 2 factors | 0.828 |
| [V$KLFS](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24KLFS) | Krueppel like transcription factors | 0.872 |
| [V$PAX6](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24PAX6) | PAX-4/PAX-6 paired domain binding sites | 0.875 |
| [V$LHXF](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24LHXF) | Lim homeodomain factors | 0.862 |
| [V$ETSF](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24ETSF) | Human and murine ETS1 factors | 0.974 |
| [V$SF1F](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24SF1F) | Vertebrate steroidogenic factor | 0.967 |
| [V$HNF1](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24HNF1) | Hepatic Nuclear Factor 1 | 0.912 |
| [V$KLFS](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24KLFS) | Krueppel like transcription factors | 0.988 |
| V$AP1R | MAF and AP1 related factors | 1 |
| [V$FKHD](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24FKHD) | Fork head domain factors | 1 |
| [V$CP2F](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24CP2F) | CP2-erythrocyte Factor related to drosophila Elf1 | 0.883 |
| [V$MTF1](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24MTF1) | Metal induced transcription factor | 0.871 |
| [V$ZF02](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24ZF02) | CP2-erythrocyte Factor related to drosophila Elf1 | 0.966 |
| [V$SPZ1](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24SPZ1) | Testis-specific bHLH-Zip transcription factors | 0.959 |
| [V$CTBP](https://www.genomatix.de/cgi-bin/matinspector_prof/matrix_help.pl?s=ae8556ee5ae600518fedd5986ebeb495;ML=113;NAME=FAM_V%24CTBP) | CT element binding protein | 0.907 |

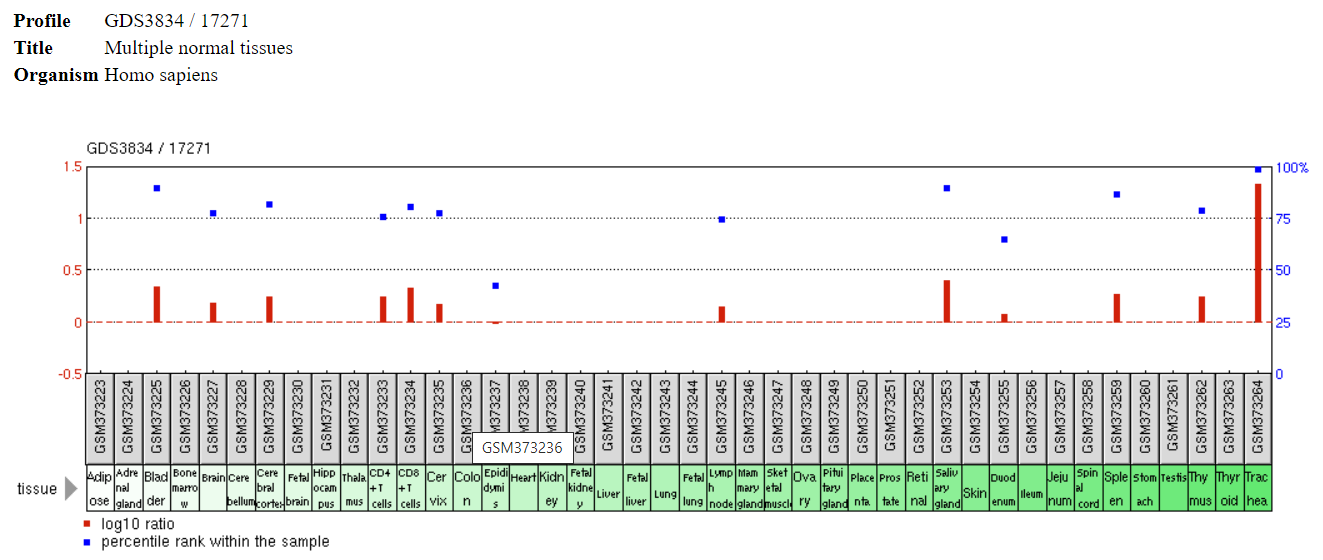
**Table 1**: Summary of the transcription factors labeled on the *ccdc190* promoter. This shows the matrix family, detailed family information and matrix similarity of all transcription factors based on information from Genomatix. Colors are corresponded to colors labeled on the promoter sequence.

To find the conservative regions for human *ccdc190* promoter, multiple sequence alignment was used. Some regions and nucleotide conservative sites are identified based on the alignment result of promoter sequences from human, horse rat and rabbit (Figure 3). These regions may indicate potential regulation or transcription bindings sites for *ccdc190* promoter.



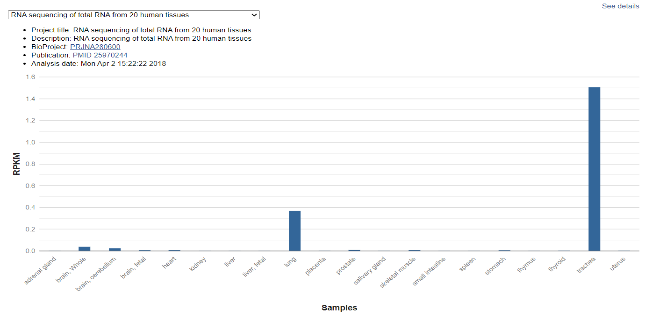
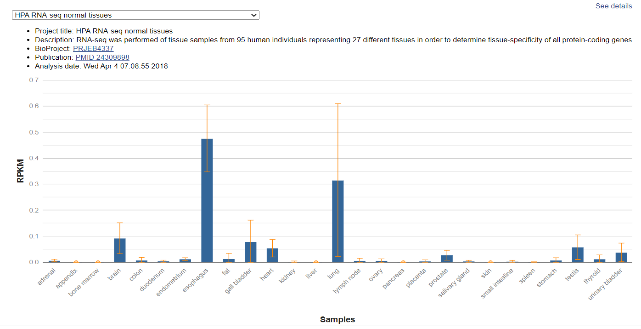
**Figure 3:** The multiple sequence alignment result of promoter regions of human, house, rat and rabbit based on Clustal. The highly conserved regions are bolded in the figure. These regions are important likely to represent conserved transcription factor binding sites.

The mRNA expression patterns in different tissues are also found in human and other species. There are some common tissues which are found to have a high expression in all four datasets like lung, brain and heart. Tissues like esophagus, trachea and stomach also have the highest expression amount in their specific measured dataset (Figure 4). The common RPKM range is from number near 0 to 1.5 (Figure 5D), or from near 0 to 0.5 (Figure 5A). Generally, *ccdc190* is identified to be highly expressed in tissues of human’s respiratory system, digestive system and reproductive system, and also found expressed in heart and brain. From figure 5D, we can also find that the expression amount of *ccdc190* tend to increase with the increase of fetal age.

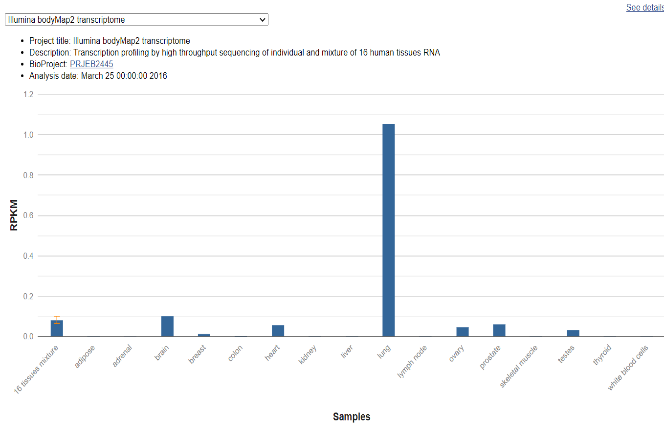
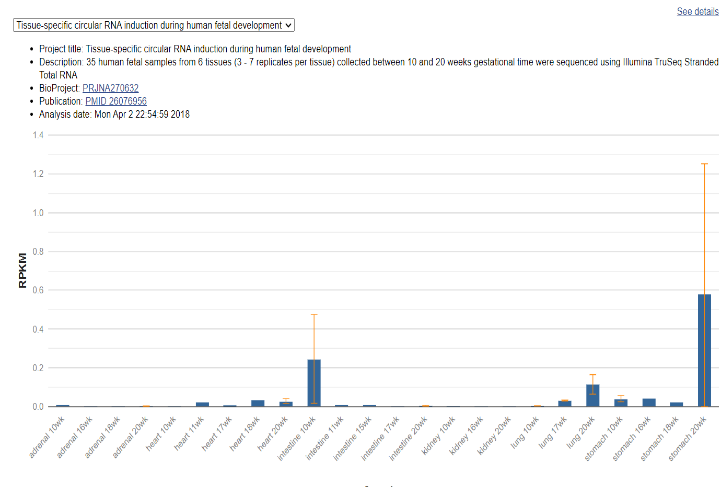


**Figure 4:** Microarray-assessed tissue expression patterns for *ccdc190* gene in human. It can be found that trachea has the highest expression amount of *ccdc190*. The expression is also identified in tissues in nervous system, digestive system and immune system. While the overall expression amount is low based on the percentage expression amount.

1. (B)

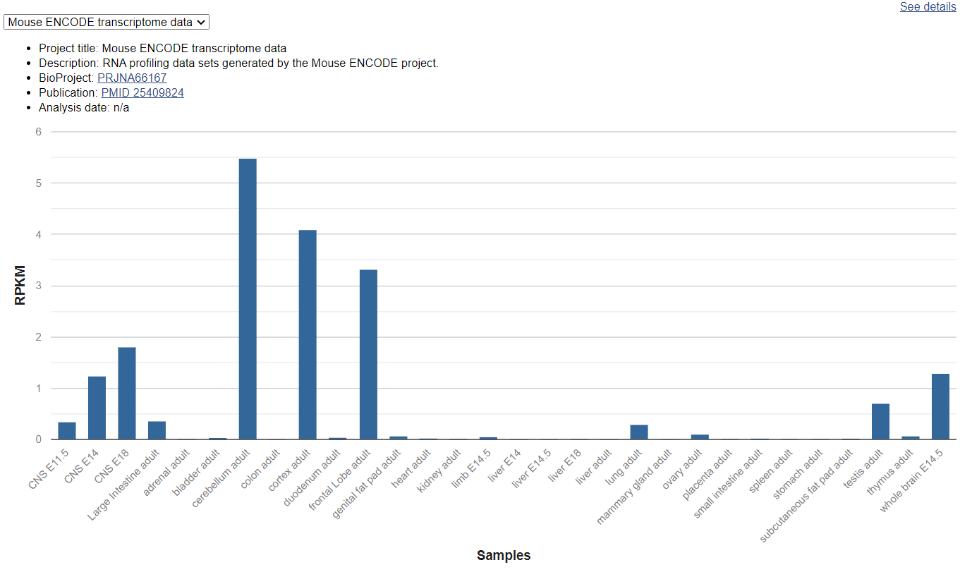
 

(C) (D)

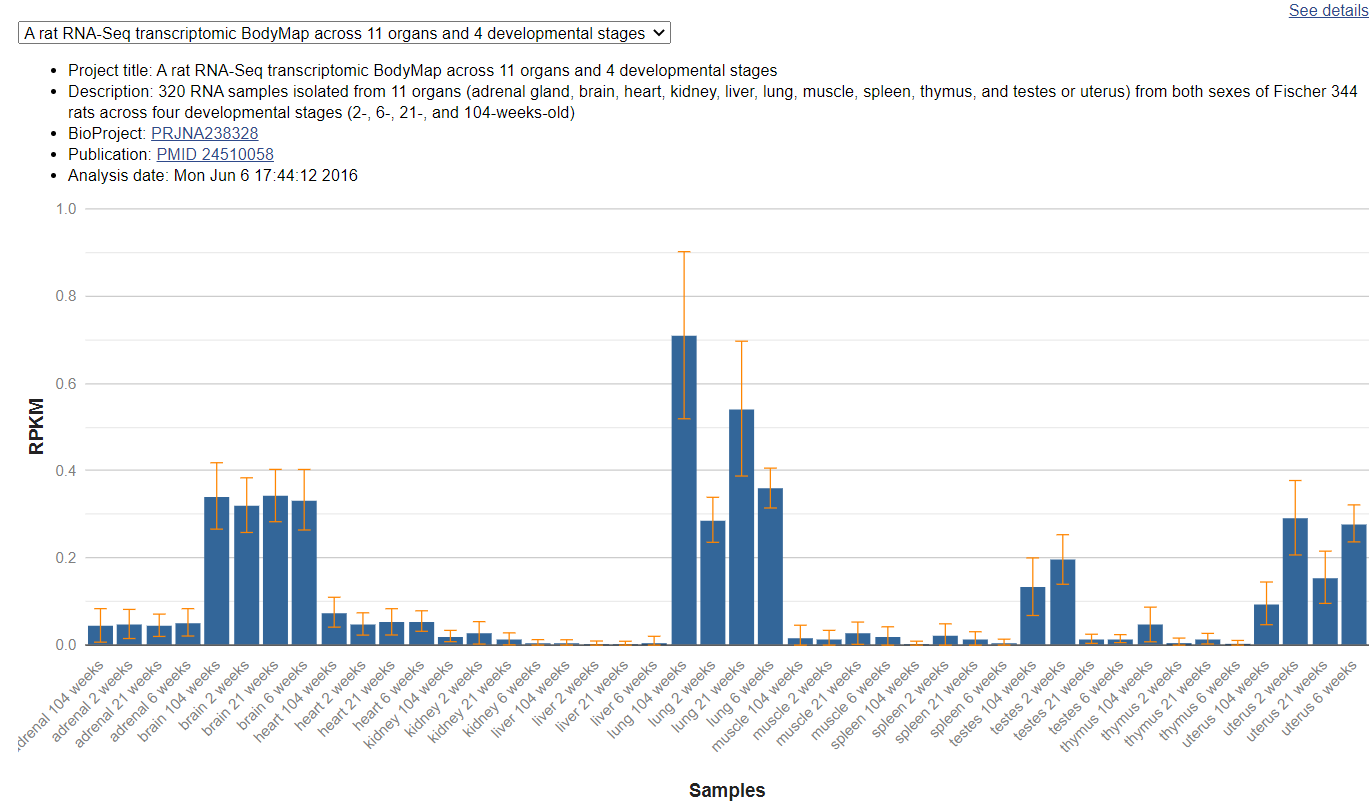
**Figure 5:** The expression pattern for human ccdc190 gene found in NCBI gene database (A-D). (A) HPA RNA-seq normal tissues for ccdc190 gene. RNA-seq was performed of tissue samples from 95 human individuals representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes. (B) RNA sequencing of total RNA from 20 human tissues. (C) The Illumina bodyMap2 transcriptome, transcription profiling by high throughput sequencing of individual and mixture of 16 human tissues RNA. (D) Tissue-specific circular RNA induction during human fetal development. 35 human fetal samples from 6 tissues (3 - 7 replicates per tissue) collected between 10- and 20-weeks gestational time were sequenced using Illumina TruSeq Stranded total RNA.

It can be found that the expression of *ccdc190* in mouse is more focused on the neural system, in which the cerebellum has the highest expression amount (Figure 6). Expression of this gene is also identified in testis, ovary and lung of mouse, though these can be five-fold lower than that of cerebellum. This is consistent with the expression pattern in human with minor differences.



**Figure 6:** Mouse ENCODE transcriptome data for ccdc190 gene. This RNA profiling data sets is generated by the Mouse ENCODE project.

Figure 7 shows the expression pattern of *ccdc190* in rat with different ages. In rat, *ccdc190* is found to be highly expressed in lung, brain, testes and uterus, and the expression amount varies with different age (Figure 7). This result is generally consistent with the patterns found in human and mice.

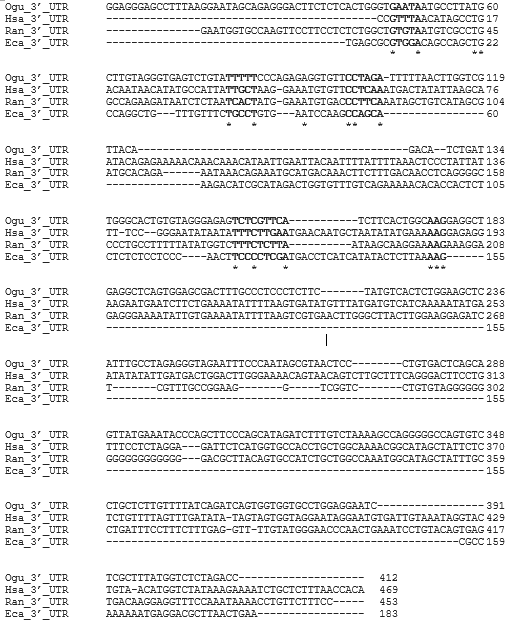
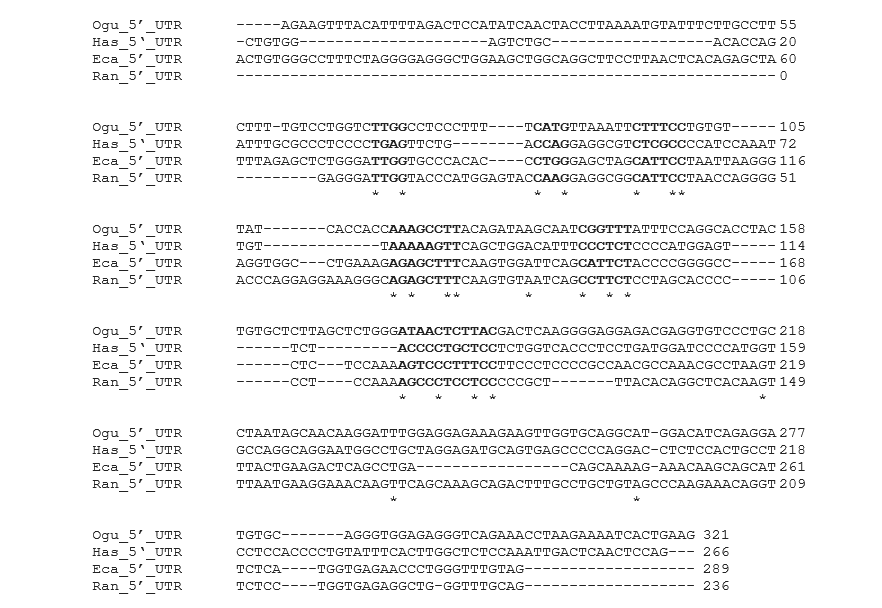


**Figure 7:**  A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. 320 RNA samples isolated from 11 organs (adrenal gland, brain, heart, kidney, liver, lung, muscle, spleen, thymus, and testes or uterus) from both sexes of Fischer 344 rats across four developmental stages (2-, 6-, 21-, and 104-weeks-old).

As conclusion, *ccdc190* is found expressed in tissues of nervous system, respiratory system, digestive system and reproductive system among human, mouse, rat and dog. While it is found highly expressed at lung and trachea of human and rat, tissues with the highest expression amount are brain related in mouse and dog. The microarray data for human also reveal that *ccdc190* is expressed in immune cell like CD4+T cells and CD8+T cells, which may suggest that there is a relation between *ccdc190* and human immune system.

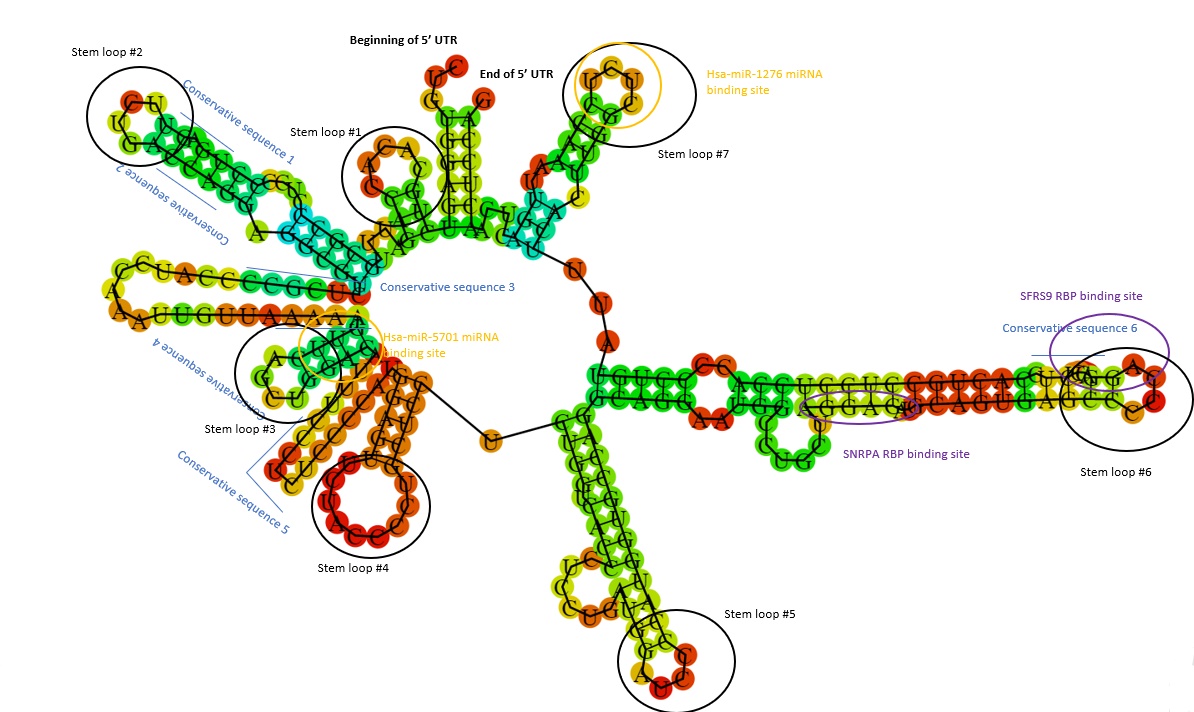
mRNA and translation regulation

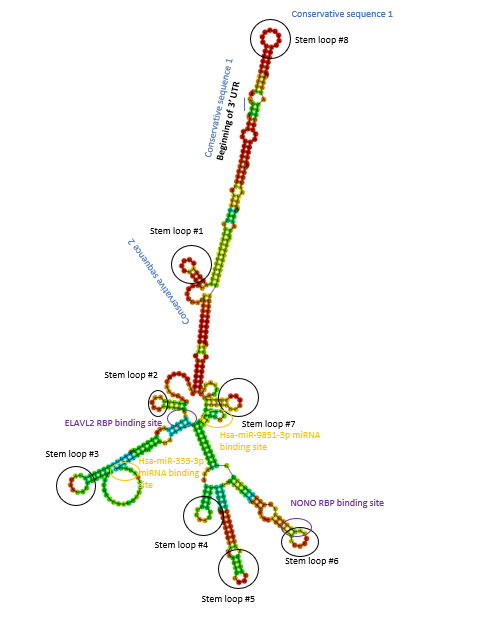
Despite the coding sequence of *ccdc190* mRNA, some conserved regions are found in the 5' UTR and 3' UTR region of human ccdc190 mRNA based on multiple sequence alignment results, they can be potential RNA binding protein and miRNA binding regions (Figure 8).



**Figure 8:** The multiple sequence alignment result of 5’ UTR and 3’ UTR regions of human, house, rat and rabbit based on Clustal. The highly conserved regions are bolded in the figure. These regions are important likely to represent conserved transcription factor binding sites.

The secondary structures of 5’ a d 3’ UTR are also predicted based on RNAFold. Some of the conservative regions, RNA binding protein and miRNA binding sites are also predicted and labelled (Figure 9).



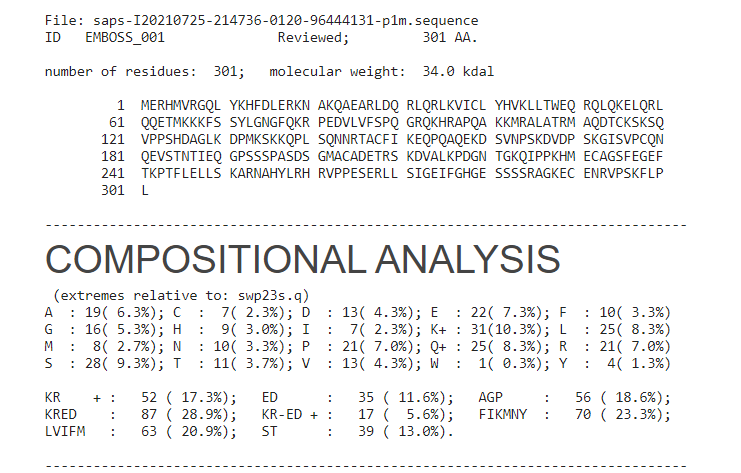


**Figure 9:** The RNA folding result of 5’ UTR and 3’ UTR regions of human based on RNAFold webset. The highly conserved regions are lined in the figure. These regions are important likely to represent conserved transcription factor binding sites. The RBP regions are identified based on information from RBPDB database, and miRNA binding sites are found from miRDB.

Protein Structure and localization

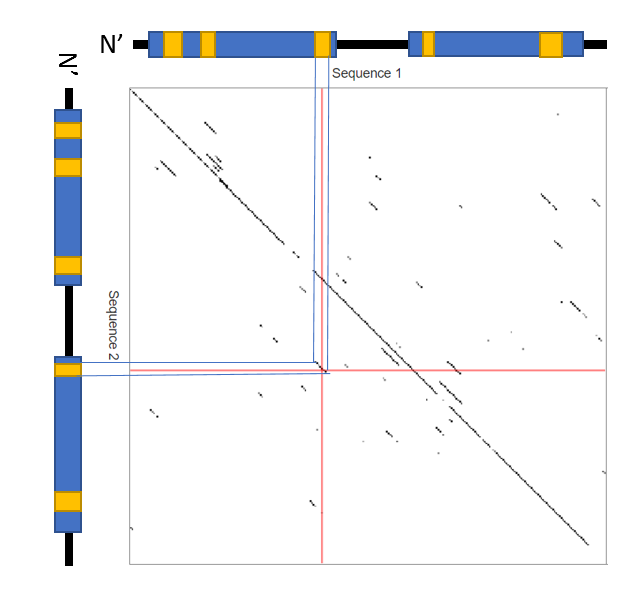
Based on the result of BLAST of human *ccdc190* protein sequence, two isoforms can be found with high similarity near 99.8%. Also, from the Protein database of NCBI it can be found that four isoforms of *ccdc*190 protein (isoform 1, isoform X1, isoform 2, isoform X3) which are derived from transcription variants.

The theoretical isoelectric point and molecular weight of human *ccdc190* isoform 1 are 9.62 and 34kD. Also, based on the analysis of SAPS, this protein seems to have more lysine and glutamine compared to other proteins in human. The sum of arginine and lysine is also higher than normal, indicating a high amount of basic amino acid in this protein. This can be also validated by the fact that the difference of basic amino acid and acidic amino acid is higher in *ccdc190* (Figure 10). The repeats are identified at the start and middle of the protein sequence. While this protein has no significant charge segments as well as hydrophobic or transmembrane domains. The similar pattern can be identified in its close orthologs from gorilla and horse. In its distant orthologs, significant charge segments and transmembrane segments are still not identified, however, the composition of amino acid can be very different with that of human, as well as the repeats pattern. 1



**Figure 10:** based on the analysis of SAPS, this protein seems to have more lysine and glutamine compared to other proteins in human. The sum of arginine and lysine is also higher than normal, indicating a high amount of basic amino acid in this protein. This can be also validated by the fact that the difference of basic amino acid and acidic amino acid is higher in *ccdc190*.

The sequence of amino acid can be explored by the result of dot matrix alignment of *ccdc190* sequence of human, with some noise processing. It shows some local alignment similarity in two sequence. It can be found that those two sequences have high global similarity due to the long nearly consistent diagonal in the graph. While the local similarities, like the region labeled by two blue lines, are relative short and sporadic. It indicates that local sequences in different regions are similar to each other. The pattern show on the figure may indicate that those similarities are more likely caused by randomness rather than particular patterns in the protein sequences.

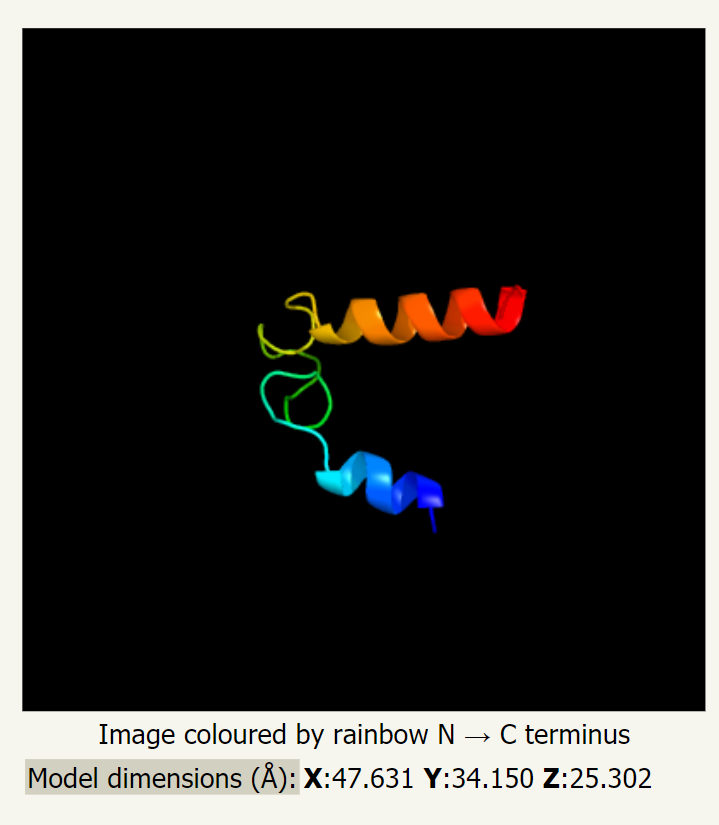


**Figure 11:** Dot matrix alignment result of human c1orf110 sequence using dotlet. Sequences show on up and left show the potential alignment pattern for two protein. Two blue lines show a region of local alignment.

Most of the secondary structure are found to be helix from prediction tools, this pattern is conserved in the close orthologs and are different in the distant orthologs (Figure 12). The predicted tertiary structure is shown on the left figure, which shows that this protein may have a helix-loop-helix motif that has the capability to interact with the DNA double strand. suggesting a potential DNA binding function of *ccdc190* (Figure 13)*.*

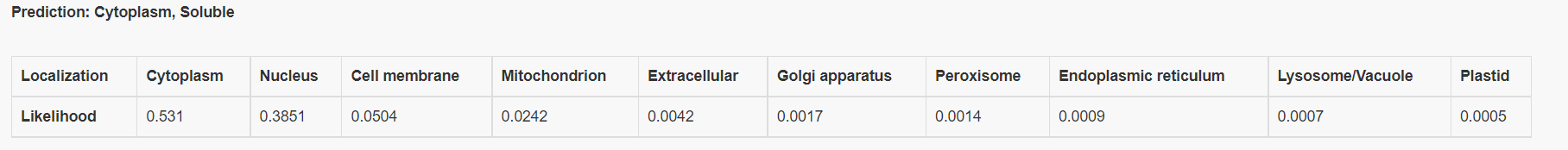


**Figure 12:** The alignment of secondary structure with the reference protein *c3zv0A*, which has 28.3% confidence to be similar to human *ccdc190* based on the prediction of Phyre2. Most of the secondary structure are helix. This pattern is conserved in the close orthologs and are different in the distant orthologs.



**Figure 13:** The predicted tertiary structure with the highest confidence of human *ccdc190* based on the results of Phyre2. Two coiled-coil domains can be identified from the figure. Regions of red color have the highest confidence and region of blue has the lowest confidence.

Based on the result of Deeploc, 53.1% of the *ccdc190* protein is in the cytoplasm, 38.5% of it is located in nucleus, others are distributed in mitochondria, endoplasmic reticulum (ER) and Golgi apparatus (Figure 14).

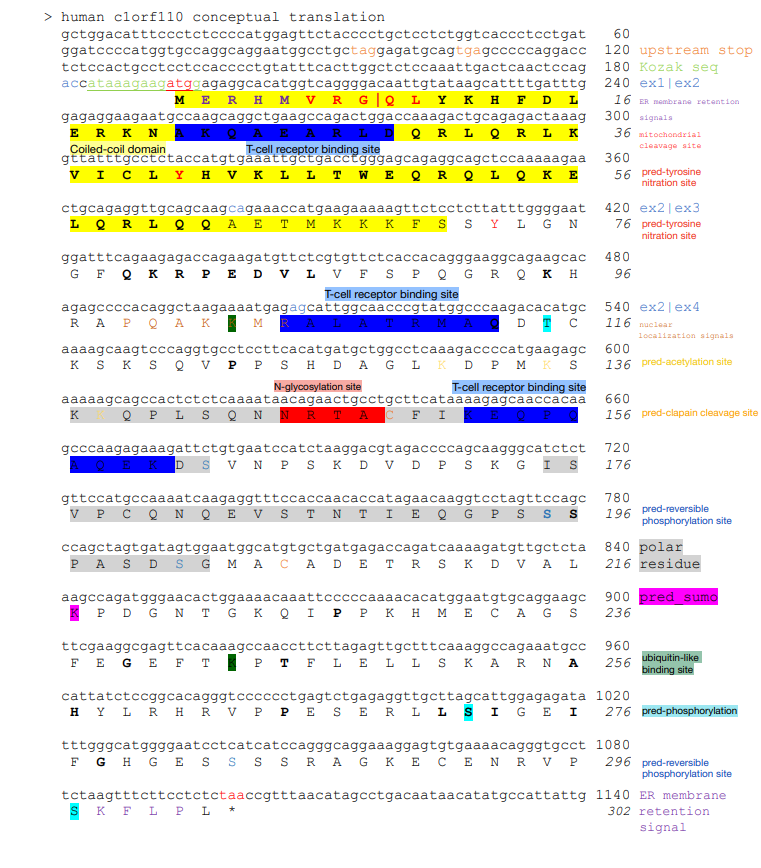


**Figure 14:** The likelihood of human *ccdc190* in different subcellular regions. This result shows that cytoplasm has the largest amount of *ccdc190* followed by nucleus, which is contradicted to the predicted result of PSORT II.

There are also predicted cleavage site for mitochondrial presequence for this protein, which makes sense due to nearly a quarter of the amount is in mitochondrial. Also, nuclear localization signals were also identified with a NLS score of 0.13. The ER membrane retention signals are also identified in the N-terminus and C-terminus. All of these are labeled on the conceptual translation in P2. There is no transmembrane region based on the predicted result. Also, no signal-peptide sequence can be identified based on the prediction of SignalP website.

Post translation modification and protein interaction

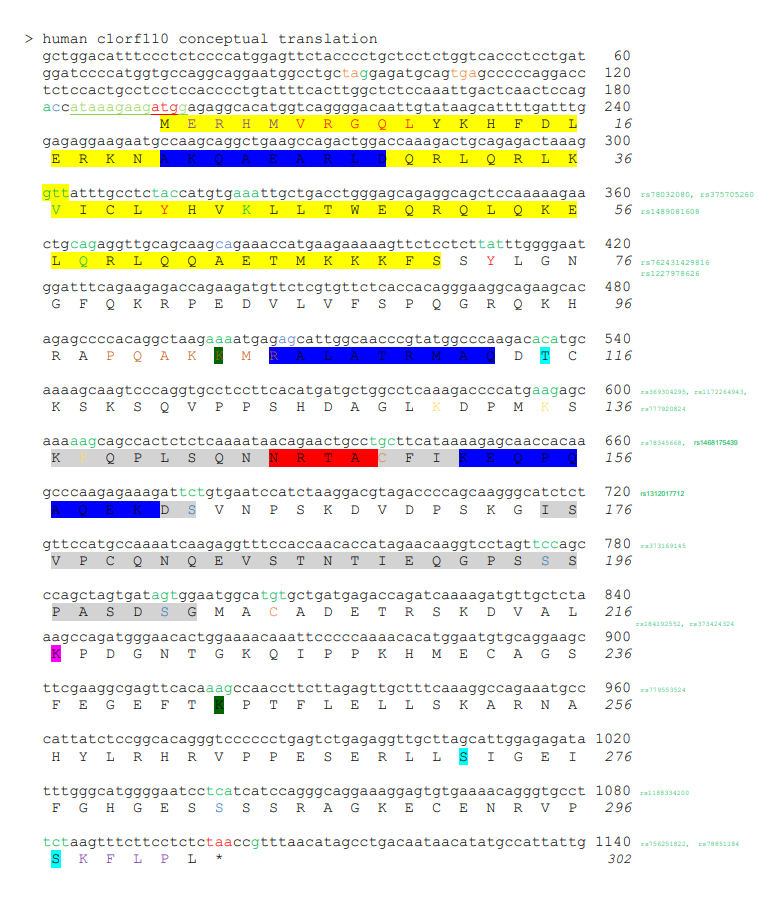
A conceptual translation of human *ccdc190* gene is made, which contained predicted domains, post translation modification sites and signal peptides (Figure 15). The predicted exon dividers, start and stop codon and conservative region of mRNA is also labelled.





**Figure 15:** The conceptual translation of human ccdc190 gene. In this figure, amino acid regions for predicted domains, post translation modification sites and signal peptides are labelled. The conserved regions found by multiple sequence alignment are bolded. The legends of each label are either shown on the right or on the margin of translation.

The single nucleotide polymorphism (SNP) is also labelled in another version of the conceptual translation (Figure 16).



**Figure 16:** The conceptual translation of human *ccdc190* gene with chosen SNP labeled, the nucleotide sites were chosen from post modification regions, conservative regions in common domains and single conserved amnio acid site. The chosen SNPs are labelled green.

|  |  |  |  |
| --- | --- | --- | --- |
| Position | Change | Significance | Cluster ID |
| mRNA 93 | a -> c | Change in upstream stop codon | rs1364260006 |
| mRNA 181 | a -> t | Change in start of Kozak sequnce | rs2032049 |
| aa 49 | V -> T | Change in most conserved region of protein | rs78032080 |
| aa 53 | Y -> A | Change in the predicted tyrosine nitration site | [rs375705260](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=375705260) |
| aa 56 | K -> A | Change in most conserved region of protein | rs1489081608 |
| aa 58 | Q -> C | Change in most conserved region of protein | [rs762431429816](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=762431429) |
| aa 85 | Y -> T | Change in the predicted tyrosine nitration site | [rs1227978626](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=1227978626) |
| aa 115 | K -> A | Change in the ubiquitin-like protein binding site | [rs369304295](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=369304295) |
| aa 127 | T -> C | Change in the predicted phosphorylation site | rs1172264943 |
| aa 146 | K -> A | Change in the potential acetylation site | rs777920824 |
| aa 150 | K -> A | Change in the potential acetylation site | rs78345668 |
| aa 161 | C -> T | Change in the calpain cleavage site | rs1468175439 |
| aa 174 | S -> C | Change in the potential reversible phosphorylation sites | rs1312017712 |
| aa 207 | S -> C | Change in the potential reversible phosphorylation sites | [rs373169145](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=373169145) |
| aa 213 | S -> G | Change in the potential reversible phosphorylation sites | [rs184192552](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=184192552) |
| aa 217 | C -> T | Change in the calpain cleavage site | rs373424324 |
| aa 255 | K -> A | Change in the ubiquitin-like protein binding site | rs779553524 |
| aa 295 | S -> C | Change in the potential reversible phosphorylation site | [rs1188334200](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=1188334200) |
| aa 297 | S -> T | Change in the predicted tyrosine nitration site | rs756251822 |
| mRNA 2004 | g -> t | Change in the conserved region of 3' UTR | [rs78851184](https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=78851184) |

**Table 2:**  A summary of chosen variant sites of human *ccdc190* sequence based on the results of SNP database of NCBI. These sites are chosen because they are within the most conserved regions or they are predicted modification sites based on result in P2. These chosen sites are labeled to green in the conceptual translation shown above.

The potential interactants of *ccdc190* is found and summarized in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Abbreviated name | Full name | Tools that reported the interaction | Basis of identification | Scores or statistical measures | Brief function |
| APP | amyloid beta (A4) precursor protein | ELISA | BioGRID |  | Proteolysis of APP generates neurotoxic Aβ peptide which is crucial for development of Alzheimer's disease |
| [NPM1](https://thebiogrid.org/110929/summary/homo-sapiens/npm1.html) | nucleophosmin | crosslinking mass spectrometry | BioGRID |  | Involved in diverse cellular processes such as ribosome biogenesis, centrosome duplication, protein chaperoning, histone assembly, cell proliferation, and regulation of tumor suppressors p53/TP53 and ARF. |
| [PPIB](https://thebiogrid.org/111475/summary/homo-sapiens/ppib.html) | peptidylprolyl isomerase B | crosslinking mass spectrometry | BioGRID |  | PPIase that catalyzes the cis-trans isomerization of proline imidic peptide bonds in oligopeptides and may therefore assist protein folding. |
| [PPP1R14C](https://thebiogrid.org/123578/summary/homo-sapiens/ppp1r14c.html) | protein phosphatase 1, regulatory (inhibitor) subunit 14C | crosslinking mass spectrometry | BioGRID |  | Inhibitor of the PP1 regulatory subunit PPP1CA. |
| PAQR6 | Membrane progestin receptor delta | Textmining | STRING | 0.721 | May be involved in regulating rapid P4 signaling in the nervous system. Also binds dehydroepiandrosterone (DHEA), pregnanolone, pregnenolone and allopregnanolone |
| C2orf66 |  | Textmining | STRING | 0.669 | Uncharaterized |
| NPSR1 | Neuropeptide S receptor | Textmining | STRING | 0.625 | Promotes mobilization of intracellular Ca(2+) stores. Inhibits cell growth in response to NPS binding. Involved in pathogenesis of asthma and other IgE-mediated diseases |
|  |  |  |  |  |  |
| EYA3 | Eyes absent homolog 3 | Textmining | STRING | 0.574 | Promotes efficient DNA repair by dephosphorylating H2AX, promoting the recruitment of DNA repair complexes containing MDC1. |
| Transmembrane protein 217 |  | Textmining | STRING | 0.556 | Uncharaterized |
| SOX-14 | Transcription factor SOX-14 | Textmining | STRING | 0.536 | Acts as a negative regulator of transcription |
| KRTAP5-2 | Keratin-associated protein 5-2 | Textmining | STRING | 0.529 | In the hair cortex, hair keratin intermediate filaments are embedded in an interfilamentous matrix, consisting of hair keratin-associated protein (KRTAP), which are essential for the formation of a rigid and resistant hair shaft through their extensive disulfide bond cross-linking with abundant cysteine residues of hair keratins. |
| TCEB3C | Elongin-A3 | Textmining | STRING | 0.526 | A general transcription elongation factor that increases the RNA polymerase II transcription elongation past template-encoded arresting sites. |

**Table 3:** The predicted interactants of human *ccdc190* collected based on the results in BioGRID and STRING. While protein reported by BioGRID do not have statistics score, they are identified from literature using experimental methods like ELISR and crosslinking spectrometry, and the results of STRING is based on textmining. Thus, the proteins reported by BioGRID is more likely to interact with *ccdc190*, along with PAQR6 which has the highest score based on result of STRING. Results from STRING can also validated using experimental methods like ELISR.

Homology and evolution

Though some isoforms are found for *ccdc190* protein in each species, no paralog can be identified for this gene. Based on the result of BLAST of human *ccdc190* protein sequence, two isoforms can be found with high similarity near 99.8%. Also, from the Protein database of NCBI we can find four isoforms of *ccdc*190 protein (isoform 1, isoform X1, isoform 2, isoform X3) which are derived from transcription variants. Thus, I conclude that no paralog can be found for *ccdc190* gene and the isoforms are derived from transcription variants due to alternative splicing.

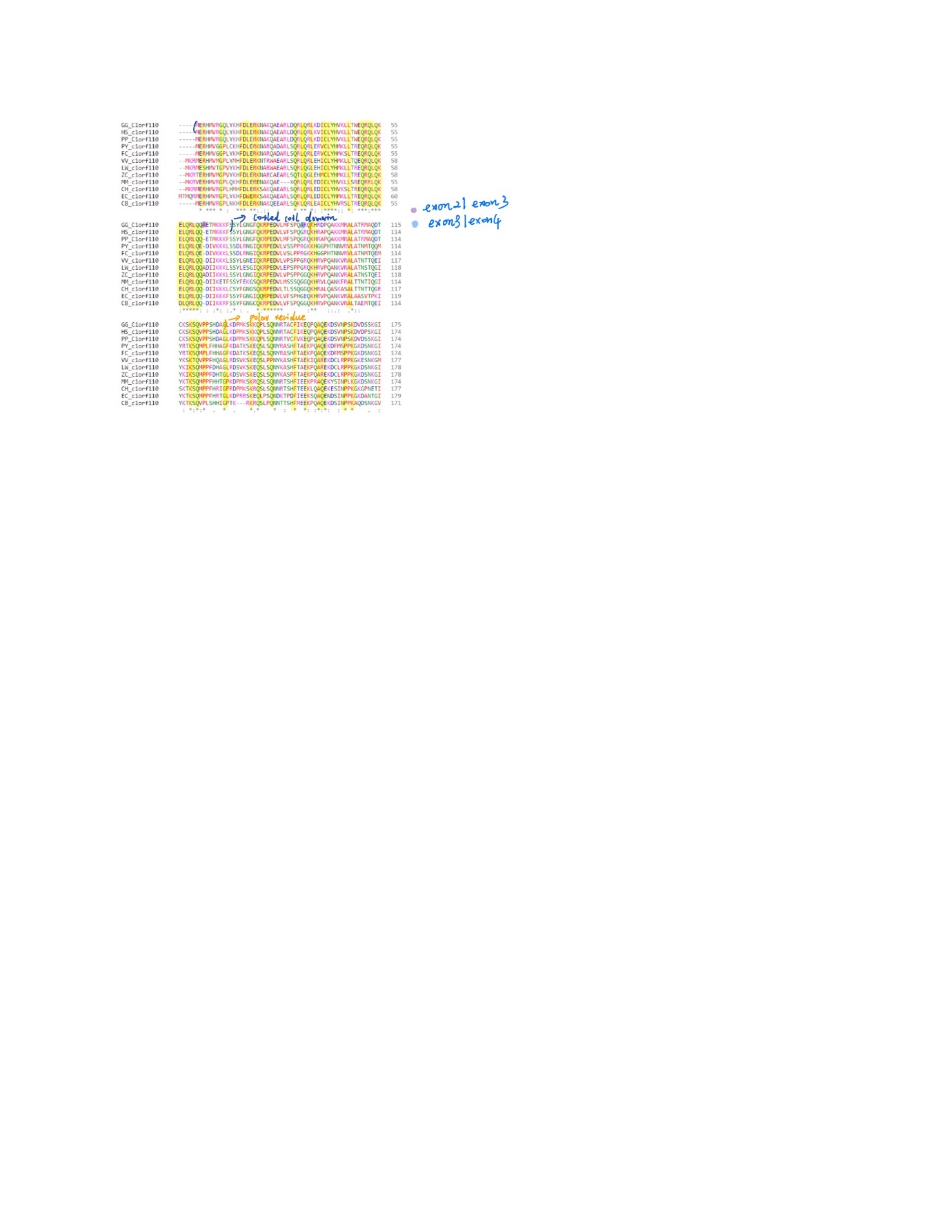
By blast, the E values for all the orthologs are from 0 to 8e-126, and the similarity is from 63.1% to 98.7% (excluding human sequence). The least related sequence is found from red fox (*Vulpes vulpes*). Then the homology of *ccdc190* is also identified in HomoloGene database of NCBI and the least relative sequence is found from coelacanth (*Latimeria chalumnae*). Thus, this gene is found to be expressed in vertebrate like fish, amphibian, reptile and mammal, expression for *ccdc190* in invertebrate is not identified.

The orthologs finding are shown in the below table, with a chosen species of 20. This dataset contains *ccdc190* sequence for 3 fishes, 2 amphibian, 3 reptile, 5 carnivora, 4 ungulate and 2 primates (Table 4). For sequences not shown in BLAST the similarity and identity are found by performing global alignment in EMBOSS Needle.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Genus and species* | Common name | Taxonomic group | Date of divergence from the human lineage (MYA) | Accession number if applicable | Sequence length (aa) | Sequence identity to human protein | Sequence similarity to human |
| *Homo sapiens* | human | [Primates](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&id=9443&lvl=3&keep=1&srchmode=1&unlock) | 0 | NP\_001380994.1 | 301 | 100 | 100 |
| *Pan paniscus* | Bonobo | [Primates](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&id=9443&lvl=3&keep=1&srchmode=1&unlock) | 6.7 | [XP\_009435065.1](https://www.ncbi.nlm.nih.gov/protein/XP_009435065.1?report=genbank&log$=prottop&blast_rank=7&RID=DMK3BV9W016) | 302 | 97.35 | 98.7 |
| *Gorilla gorilla* | Western gorilla | [Primates](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&id=9443&lvl=3&keep=1&srchmode=1&unlock) | 9.06 | [XP\_004027846.3](https://www.ncbi.nlm.nih.gov/protein/XP_004027846.3?report=genbank&log$=prottop&blast_rank=8&RID=DMK3BV9W016) | 302 | 97.4 | 97.7 |
| *Puma yagouaroundi* | Jaguarundi | Carnivora | 96 | XP\_040316822.1 | 331 | 57.4 | 66.7 |
| *Felis catus* | Cat | Carnivora | 96 | XP\_019677815.1 | 331 | 57.4 | 66.7 |
| *Vulpes vulpes* | red fox | Carnivora | 96 | [XP\_025847051.1](https://www.ncbi.nlm.nih.gov/protein/XP_025847051.1?report=genbank&log$=protalign&blast_rank=95&RID=DMK3BV9W016) | 302 | 62.8 | 74.3 |
| *Zalophus californianus* | California sea lion | Carnivora | 96 | XP\_027467744.1 | 302 | 63 | 71.8 |
| *Capra hircus* | Goat | Artiodactyla | 96 | [XP\_005677176.1](https://www.ncbi.nlm.nih.gov/protein/XP_005677176.1?report=genbank&log$=protalign&blast_rank=90&RID=DMK3BV9W016) | 299 | 63.2 | 71.7 |
| *Monodon monoceros* | narwhal | Artiodactyla | 96 | [XP\_029082760.1](https://www.ncbi.nlm.nih.gov/protein/XP_029082760.1?report=genbank&log$=prottop&blast_rank=89&RID=DMK3BV9W016) | 296 | 63.5 | 74 |
| *Leptonychotes weddellii* | Weddell seal | Carnivora | 96 | XP\_006734577.2 | 303 | 63.6 | 73.4 |
| *Camelus bactrianus* | Bactrian camel | Artiodactyla | 96 | [XP\_010952087.1](https://www.ncbi.nlm.nih.gov/protein/XP_010952087.1?report=genbank&log$=protalign&blast_rank=72&RID=DMK3BV9W016) | 279 | 64.1 | 74.8 |
| *Equus caballus* | house | Perissodactyla | 96 | [XP\_014595175.1](https://www.ncbi.nlm.nih.gov/protein/XP_014595175.1?report=genbank&log$=protalign&blast_rank=62&RID=DMK3BV9W016) | 287 | 64.4 | 73.5 |
| *Alligator mississippiensis* | American alligator | Archosauria | 312 | XP\_014460644.2 | 367 | 22.8 | 34.4 |
| *Zootoca vivipara* | common lizard | Lepidosauria | 312 | XP\_034979493.1 | 293 | 24 | 37.8 |
| *Chelonia mydas* | Green sea turtle | Testudinata | 312 | XP\_027674491.2 | 309 | 28.4 | 42.2 |
| *Bufo bufo* | common toad | Batrachia | 351.7 | XP\_040264705.1 | 263 | 23 | 39 |
| *Rana temporaria* | common frog | Batrachia | 351.8 | XP\_040215666.1 | 287 | 23.8 | 39.9 |
| *Latimeria chalumnae* | coelacanth | Coelacanthimorpha | 413 | XP\_006014086.1 | 348 | 24.5 | 35.8 |
| *Lepisosteus oculatus* | spotted gar | Actinopteri | 435 | XP\_015212080.1 | 236 | 26.7 | 38.9 |
| *Callorhinchus milii* | elephant shark | Gnathostomata | 473 | XP\_007893833.1 | 337 | 24.7 | 40.8 |

**Table 4:** Twenty species found that have orthologs with human *ccdc190* protein. Species that are in same group are labeled with same color. The Taxonomic group column is filled based on Taxonomy database of NCBI. Date of divergence from the human is found in time tree website. Sequence identity or similarity to human are either found from BLAST result or from output of Needle alignment.

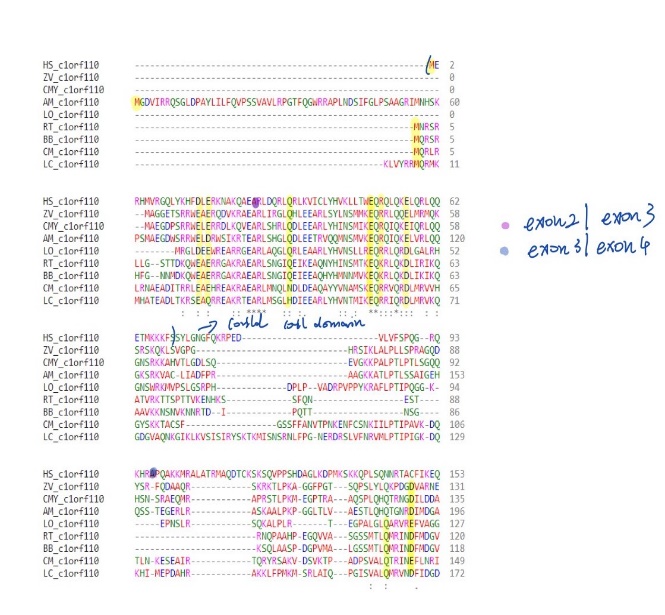
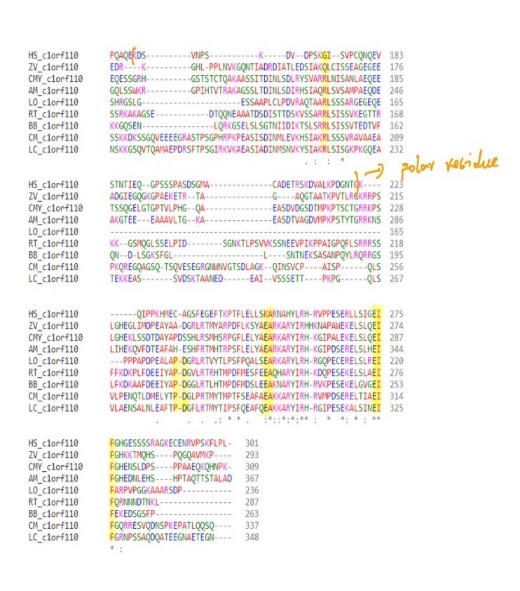
Based on the output of clustal, the strict orthologs to human are: *Gorilla gorilla, Pan paniscus, Zalophus californianus, Puma yagouaroundi, Felis catus, Vulpes vulpes, Leptonychotes weddellii, Monodon monoceros, Equus caballus, Camelus bactrianus* and *Capra hircus.* Then protein sequences from these species are clustered to find the conservative regions (Figure 17).



**Figure 17:** The multiple alignment of strict orthologs to human *ccdc190* proteins. Conserved regions are yellow labelled. The coiled coil domain and polar residues are also labeled based on information of Uniprot, along with the exon dividers.

It can be found that in the strict orthologs to human ccdc190 proteins, there are many conservative amnio acids showing a high similarity between those sequences. For example, Glu2 is conserved in all non-human primates and some carnivores.

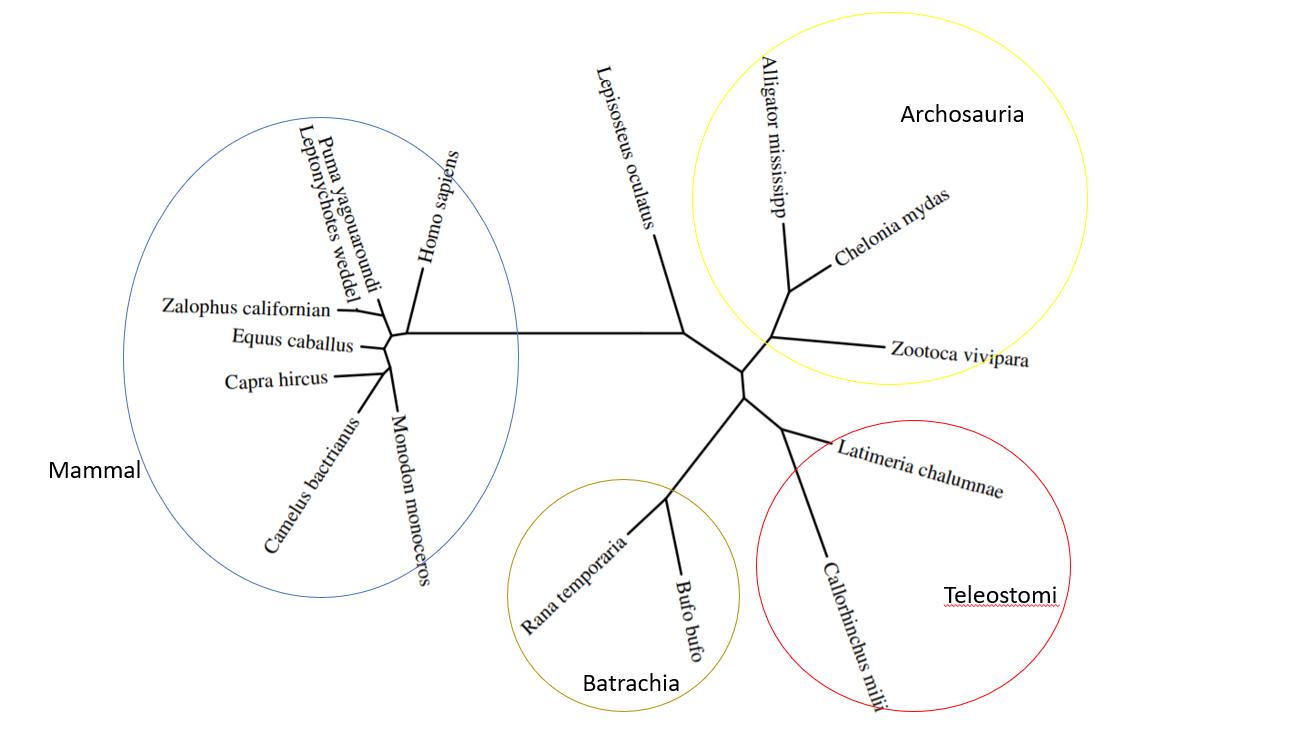
The distant homologs are: *Rana temporaria, Bufo bufo, Zootoca vivipara, Chelonia mydas, Alligator mississippiensis, Lepisosteus oculatus, Callorhinchus milii* and *Latimeria chalumnae*. Then protein sequences from these species are clustered to find the conservative regions (Figure 18).

**Figure 18:** The multiple alignment of distant orthologs to human *ccdc190* proteins. Conserved regions are yellow labelled. The coiled coil domain and polar residues are also labeled based on information of Uniprot, along with the exon dividers.

Common residues in distant orthologs are significantly fewer than that of strict orthologs. Still, there are some highly conserved amino acids like Glu49 that is conserved in human, reptile, amphibian and fish for which sequences are available.

Based on the analysis in table 4, I found that this gene first appeared in the elephant shark (*Callorhinchus milii*), which is 473 million years ago from now. While the gene family cannot be determined due to the lack of paralogs. The most distantly related organism is found to be coelacanth (*Latimeria chalumnae*) with a sequence identity of 24.8%. The number of alternative splicing variants in it can be hard to determine due to the lack of data. While we do know its mRNA has 3 exons which can be used to infer the theoretical number of the variants.



**Figure 19:** The time-calibrated unrooted phylogenetic tree for *ccdc190* protein. Most species are clustered to different groups. *Archosauria* contains reptiles, *Teleostomi* contains fish, and *Batrachia* contains amphibian. Species in mammal group are tightly close to each other. *Lepisosteus oculatus*is an outlier since only partial sequence is acquired in that species.

It is also interested to find how fast this protein evolve compared to others. To achieve that, I try to plot the divergence rate of *ccdc190* and calibrated it with two known protein (Figure 20). A relative high divergence and evolution rate was found from the plot.

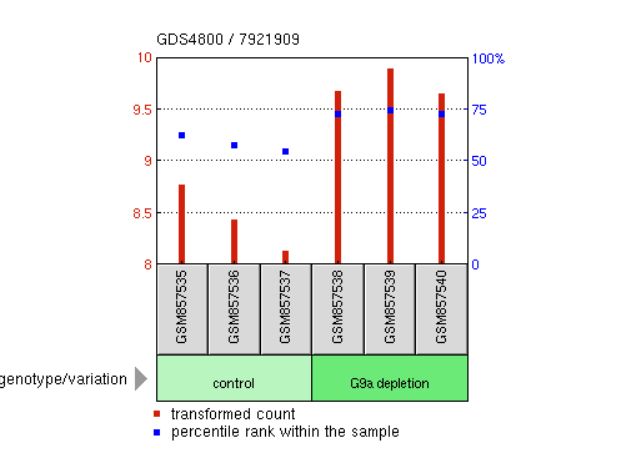
**Figure 20:** The graph showing the approximate date of divergence (from human) for a given species (MYA [million years ago]) versus the corrected % divergence (m) of that species' orthologous protein. The data and trendlines of cytochrome C and fibrinogen alpha chain are also included as indicators for proteins with low divergence and high divergence. As it can be found from the graph, the trendline of *ccdc190* is closer to that of fibrinogen alpha, showing a relative high divergence and evolution rate.

Function and clinical significance

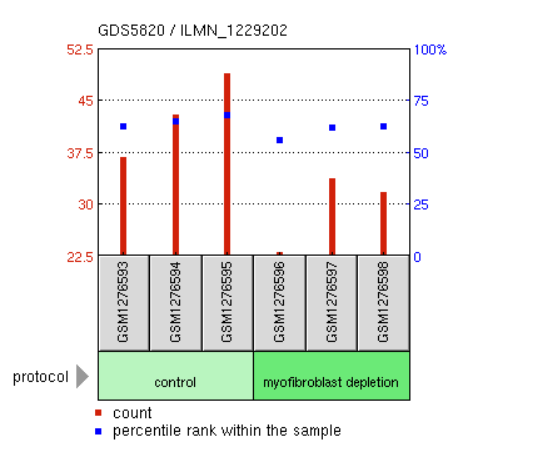
The human *ccdc190* protein is predicted to have [coiled-coil domain](https://en.wikipedia.org/wiki/Coiled_coil) that may have multiple functions. First. the physical properties like length and flexibility of this domain make it function as a good molecular spacer. It is found that the length variation of coiled-coil domain is 3.6 times lower than that of other regions, and the length of amino acid is relatively conserved within different proteins, indicating that the physical size of this domain may play an important role in its function.1 Also, coiled-coil domain is found essential in accurate chromosome segregation. It is found that during cell division, some coiled-coil contain proteins like [Ska1](https://www.uniprot.org/uniprot/Q96BD8), [Ska2](https://en.wikipedia.org/wiki/SKA2), and [Ska3](https://www.uniprot.org/uniprot/Q8IX90) form a dimer and the mutation of the dimer leads to chromosome congression failure which can lead to cell death.2 Finally, there is evidence that this domain can play a role in DNA recognition and binding.3 In the [restriction enzyme](https://en.wikipedia.org/wiki/Restriction_enzyme) found inside [*Methanococcus jannaschii*](https://en.wikipedia.org/wiki/Methanocaldococcus_jannaschii)*,* subunits that encoding coiled-coil domain are found served as molecular ruler for the recognition of direct repeats in DNA sequence, similar pattern is found in the [MerR](https://www.uniprot.org/uniprot/P22853) family of transcriptional activators in bacteria.[[2]](#footnote-2)

The gene *ccdc190* is found to be related to some diseases. For example, it is found contributed significantly to [cannabis dependence](https://en.wikipedia.org/wiki/Cannabis_dependence) risk by sequence kernel association tests and found to be downregulated in [nasopharyngeal carcinoma](https://en.wikipedia.org/wiki/Nasopharyngeal_carcinoma) microarray data.4,5 Also, this gene is found to be a biomarker for short photoperiods for cells in [pars tuberalis](https://en.wikipedia.org/wiki/Pars_tuberalis).6 The detailed relation for *ccdc190* with phenotypes mentioned above still needs further study.

Based on the [microarray](https://en.wikipedia.org/wiki/Microarray) data on GEO database of NCBI, it can be found that the H3K9me2 methyltransferase G9a depletion can increase the expression amount of c1orf110 on breast cancer cell line. indicating that DNA methylation caused by breast cancer can lead to inhibition of ccdc190 expression (Figure 21). Also, the αSMA+ myofibroblast early depletion on [pancreatic ductal adenocarcinoma](https://en.wikipedia.org/wiki/Pancreatic_cancer) mouse can lead to lower expression of ccdc190 in pancreas, suggesting the potential function of this gene in pancreatic cancer pathology (Figure 22).



**Figure 21:** Microarray data for *ccdc190* expression amount of both G9a depletion and control group which found from Geo database of NCBI. The data was analyzed using t-test in R with a p-value of 0.01.



**Figure 22:** Microarray data for *ccdc190* expression amount of both myofibroblast depletion and control group which found from Geo database of NCBI. The data was analyzed using t-test in R with a p-value of 0.04.

1. 1 GeneCards (GeneCards, The Human Gene Database) entry on CCDC190 [https://www.genecards.org/cgi-bin/carddisp.pl?gene=CCDC190].

   2 NCBI gene (National Center for Biotechnology information gene) entry on CCDC190 [https://www.ncbi.nlm.nih.gov/gene/339512].

   3 UniProt (UniProt Knowledgebase) entry on Q86UF4 [https://www.uniprot.org/uniprot/Q86UF4].

   4 AceView (National Center for Biotechnology information AceView) entry on c1orf110 [https://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=human&term=C1orf110&submit=Go]. [↑](#footnote-ref-1)
2. 1 Truebestein L, Leonard TA (September 2016). ["Coiled-coils: The long and short of it"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5082667). *BioEssays*. **38**(9): 903–16. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1002/bies.201600062](https://doi.org/10.1002%2Fbies.201600062). [PMC](https://en.wikipedia.org/wiki/PMC_(identifier)) [5082667](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5082667). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [27492088](https://pubmed.ncbi.nlm.nih.gov/27492088).

   Surkont J, Diekmann Y, Ryder PV, Pereira-Leal JB (December 2015). "Coiled-coil length: Size does matter". *Proteins*. **83** (12): 2162–9. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1002/prot.24932](https://doi.org/10.1002%2Fprot.24932). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [26387794](https://pubmed.ncbi.nlm.nih.gov/26387794).

   3Jeyaprakash AA, Santamaria A, Jayachandran U, Chan YW, Benda C, Nigg EA, Conti E (May 2012). "Structural and functional organization of the Ska complex, a key component of the kinetochore-microtubule interface". *Molecular Cell*. **46** (3): 274–86. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1016/j.molcel.2012.03.005](https://doi.org/10.1016%2Fj.molcel.2012.03.005). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [22483620](https://pubmed.ncbi.nlm.nih.gov/22483620).

   4 Gizer IR, Bizon C, Gilder DA, Ehlers CL, Wilhelmsen KC (January 2018). ["Whole genome sequence study of cannabis dependence in two independent cohorts"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5522771). Addiction Biology. 23 (1): 461 - 473. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1111/adb.12489](https://doi.org/10.1111%2Fadb.12489). [PMC](https://en.wikipedia.org/wiki/PMC_(identifier)) [5522771](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5522771). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [28111843](https://pubmed.ncbi.nlm.nih.gov/28111843)

   5 Ye Z, Wang F, Yan F, Wang L, Li B, Liu T, et al. (April 2019). ["Bioinformatic identification of candidate biomarkers and related transcription factors in nasopharyngeal carcinoma"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6444505). *World Journal of Surgical Oncology*. **17** (1): 60. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1186/s12957-019-1605-9](https://doi.org/10.1186%2Fs12957-019-1605-9). [PMC](https://en.wikipedia.org/wiki/PMC_(identifier)) [6444505](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6444505). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [30935420](https://pubmed.ncbi.nlm.nih.gov/30935420).

   6 Wood SH, Christian HC, Miedzinska K, Saer BR, Johnson M, Paton B, et al. (October 2015). "Binary Switching of Calendar Cells in the Pituitary Defines the Phase of the Circannual Cycle in Mammals". *Current Biology*. **25** (20): 2651–62. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1016/j.cub.2015.09.014](https://doi.org/10.1016%2Fj.cub.2015.09.014). [PMID](https://en.wikipedia.org/wiki/PMID_(identifier)) [26412130](https://pubmed.ncbi.nlm.nih.gov/26412130). [↑](#footnote-ref-2)