Phylogenetic Comparison and Splicing Analysis of U1 snRNP-specific Protein U1C in Eukaryotes

Mo-Xian Chen (cmx2009920734@gmail.com)
Chinese University of Hong Kong

Kai-Lu Zhang
Nanjing Forestry University college of biology and the environment

Jian-Li Zhou
Shenzhen Children's Hospital

Jing-Fang Yang
Central China Normal University

Yu-Zhen Zhao
Shenzhen Children's Hospital

Das Debatosh
The Chinese University of Hong Kong

Ge-Fei Hao
central china normal university

Cai Wu
Nanjing Forestry University College of Light Industry and Food Engineering

Jianhua Zhang
Hong Kong Baptist University Department of biology

Fu-Yuan Zhu
nanjing forestry university college of biology and the environment

Shao-Ming Zhou
Shenzhen Children's Hospital

Research article

Keywords: alternative splicing, phylogenetics, gene expression, splice site, U1-snRNP

DOI: https://doi.org/10.21203/rs.3.rs-142800/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

As a pivotal regulator of 5’ splice site recognition, the U1 small nuclear ribonucleoprotein (U1 snRNP)-specific protein C (U1C) regulates pre-mRNA splicing by interacting with other components of U1 snRNP complex. Previous studies have shown that U1 snRNP and its components are linked to variety of diseases including cancer. However, phylogenetic relationship and expression profiles of U1C have not been studied systematically. To this end, we identified a total of 110 animal U1C genes and compared to the homologs from yeast and plants. Bioinformatics analysis shows that the structure and function of U1C proteins is relatively conserved, and is found in multiple copies in few members of U1C genes family. Furthermore, expression pattern showed that U1Cs have potential role in cancer progression and human development. In summary, our study presents a comprehensive overview of animal U1C gene family, which can provide fundamental data and potential cues for further research in deciphering the molecular function of this splicing regulator.

Introduction

Discovered in the late 1970s, precursor message RNA (pre-mRNA) splicing including constitutive splicing and alternative splicing is an important biological process in eukaryotes [1, 2]. This process is performed by a large protein complex called spliceosome which removes the noncoding sequences (introns) and ligates functional coding sequences (exons) to generate mature mRNAs [3]. The spliceosomes are composed of multiple proteins and several U-rich small nuclear ribonucleoproteins (snRNPs) including U1, U2, U4/U6, U5, U11 and U12 [4]. Splicing machinery is assembled in a stepwise manner by different spliceosomal snRNPs [5]. The U1 snRNP-mediated recognition of downstream 5’ splice site of introns is the first step by a subcomplex during spliceosome assembly, which is composed of one U1 snRNA, 8–9 Sm proteins and three specific proteins including U1-70K, U1A and U1C in human and yeast [5, 6]. U1-70K and U1A can bind directly to the U1 snRNA while U1C alone cannot attach to U1 snRNA and the binding of U1 snRNP core domain and U1C needs to be mediated by U1-70K and Sm proteins [7, 8]. Importantly, U1C protein is specifically associated with the U1 snRNP and plays a peculiar role in recognizing the 5’ splice site independent of the base pairing [7, 9].

Biological functions of U1 snRNP have been characterized to affect human pathogenesis and autoimmune diseases besides its 5’ SS recognition. For instance, pathological examination, proteomics and transcriptomics have demonstrated that U1 snRNP components aggregate in neuronal cell bodies, resulting in a global disruption of RNA processing in Alzheimer’s disease (AD) brains [10]. Furthermore, U1 snRNP is demonstrated to be linked to a molecular pathway associated with amyotrophic lateral sclerosis (ALS), indicating that splicing defects may play a key role in the pathogenesis of motor neuron disease, providing a potential therapeutic target [11]. In addition, U1 snRNP is also reported to associate with other diseases such as autoimmunity connective tissue disease (MCTD), systemic lupus erythematosus congenital myasthenic syndrome (CMS) and others [12–15]. All these research works demonstrate that U1 snRNPs play an important role in human disease. As a vital component of the U1 snRNP, the phylogeny and splicing pattern of U1C remains unclear. Thus in this study, we identified and
analyzed the phylogenetic relationships of *U1C* gene family in different animal species. Genome-wide bioinformatics analysis including elucidation of gene structures, protein domains, expression profiles and conserved splicing patterns has been performed to explore the potential functions of *U1Cs*, providing theoretical support for further functional investigation.

**Experimental Methods**

**Sequence identification and collection of U1C proteins in animal**

The U1C protein sequence (ENSP00000363129.3) from *Homo sapiens* was used as a query sequence to perform protein BLAST (e-value cutoff = 1e⁻¹⁰) to find similar sequences in all the available animal and yeast genomes (present in Ensembl genome browser 96 (http://asia.ensembl.org/index.html)) [16]. The resulting sequences were screened for PF06220 (U1 zinc finger, zf-U1) domain using the HMMSCAN algorithm implemented in HMMER 3.2.1 [17], after which 110 protein sequences were retained.

**Construction of phylogenetic tree of U1C genes**

Above obtained 110 proteins were used for multiple sequence alignment in Muscle v3.8 [18] and for construction of phylogenetic tree of animal *U1C* genes using PhyML v3.037 based on the maximum likelihood method with JTT + G + F model [19]. Meanwhile, the sequences of plants *U1C* genes derived from our previous work were combined to build up the tree of plants, yeast, and animal with FastTree (ref: FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments). The resulting phylogenetic trees were visualized in FigTree v1.4.3.38 [20].

**Analysis of gene structures, protein domains and MEME motifs**

Genomic, cDNA, CDS and peptide sequences for above identified U1Cs were downloaded from the Ensembl database. Gene structure was reconstructed on Gene Structure Display Server 2.0 (GSDS2.0) (http://gsds.cbi.pku.edu.cn) [21]. Protein domains were found using HMMER website (https://www.ebi.ac.uk/Tools/ hmmer/) [22] and was drawn using TBtools [23]. cDNA and protein sequences were used as an input to find 10 most enriched motifs on MEME (multiple Em for motif elicitation) server operated with the default parameters (http://meme-suite.org/tools/meme) [24].

**Analysis of protein interaction networks**

Protein sequences of human (ENSP00000363129.3), *Mus musculus* (ENSMUSP00000156644.1) and *Saccharomyces cerevisiae* (YLR298C_mRNA) were used as input to obtain the protein-protein interaction networks on STRING web server (https://string-db.org/) [25] using its active interaction sources: experiments and databases.
Amino acid conservation estimation

The crystal structure of human U1C protein (PDB ID: 4PJO) downloaded from PDB database [26] was used as an input file in the ConSurf server to represent the evolutionary conservation of animal U1C genes. The corresponding model of plant was retrieved from a previous work. The residue numbers of human U1C were labeled according to the human sequence (ENSP00000363129.3) in the representative figure. The sequences with large gaps were deleted in this analysis.

AS profile analysis and identification of conserved splice sites

Sequences of all available splice isoforms of animal U1C genes were downloaded from the Ensembl database. Selected splice junction sequences (20 bp on each side) were aligned using online webtool MultAlin (http://multalin.toulouse.inra.fr/multalin/ multalin.html) and graphed in Weblogo v3.0 (https://weblogo.berkeley.edu/logo.cgi).

Expression analysis from online microarray datasets

Expression data of human and mouse U1Cs were downloaded from the Expression Atlas database (https://www.ebi.ac.uk/gxa/home). The heatmaps of retrieved expression data were generated as described previously [23].

Results

Phylogenetic analysis of animal U1C genes

In this study, a total of 110 U1C protein sequences from 61 animal species, including 93 placentals (51 primates, 28 rodents and lagomorphs and 14 other mammals), 5 marsupials, monotremes and reptiles, one other vertebrate (Xenopus tropicalis), 7 fish, and 4 other species were identified (Table S1). More than half of the members of the U1C family involved multiple gene copies (72/110), including twelve species with two copies, seven species with three copies and five species with four copies. Particularly, it was found that there were seven copies in Ma’s night monkey (Aotus nancymaeae). Finally, only 38 animal species contained one copy of U1C.

To understand the evolutionary history and phylogenetic relationship between the above-identified 110 U1C genes, a rooted circle phylogenetic tree of the family constructed based on multiple protein sequences alignment (Fig. 1), showed four major clades, namely placentals (purple), marsupials, monotremes and reptiles (pink), fish (light blue) and other species (yellow). Furthermore, five members of yellow clade (other species) with longer branch length formed the basal part of the circle phylogenetic tree, suggesting its distant association with other clades. More specifically, it was observed that xenopus (Xenopus tropicalis, ENSXETP00000053216.1) gathers to one branch point with other vertebrates (placentals, marsupials, monotremes, reptiles and xenopus), which suggests that lamprey is a significant
link which connects vertebrates and invertebrates (Fig. 1). Moreover, placentals, marsupials, monotremes and reptiles formed a sister clade with xenopus (Xenopus tropicalis, ENSXETP00000053216.1).

**Protein domain and conserved motifs analysis**

Protein domain and conserved motifs were analyzed to further infer the functional properties of U1C proteins. Most of the U1Cs contained a domain called zf-U1 (PF06220) as predicted by HMMER website with the exception of three proteins (ENSPTRP00000071198.1, ENSMNEP00000066881.1 and ENSJJAP00000018575.1) without any signatures in their sequence (Fig. S2, middle panel). Identified U1C proteins from all species were characterized to range from 99 to 231 amino acids in length (average length 163.2 aa) (Table S2). 10 most conserved motifs in animal U1C proteins, analyzed by MEME suite (Fig. S2, right panel), showed that only 16 sequences of U1Cs including human U1C have ten conserved motifs identified in this work, indicating the divergence of animal U1C proteins in terms of conserved domain content. Barring one sequence ENSHHAP00000001057.1 (Sarcophilus harrisii) from marsupials, other 15 sequences were from placentals. The zf-U1 domain was present in the first three motifs at the N-terminal of most U1Cs (Fig. S2, right panel). Furthermore, most U1Cs from placentals contain nine conserved motifs, while those from marsupials, monotremes and reptiles have around eight or nine motifs and U1Cs of fish had around seven or eight motifs (Fig. S2, right panel). Expectedly, protein sequences from other species in the basal part of the phylogenetic tree had only one to five motifs and showed least degree of conservation.

**Conservation analysis and interaction networks of U1C**

Since the crystal structure of the human U1C (PDB ID: 4PJO) RRM domain is publicly available, the domain evolutionary conservation analysis was based on this structure (the residue number of human U1C as shown below). The ConSurf Grade of 20 (39.2%) residues are over 7 and the ConSurf Grade of 10 (19.6%) residues are over 9. Meanwhile, the conservations of 50 amino acids are more than 90% among 51 sites, indicating the high conservation of this gene in the animal.

In detail, His45 and His51 in the zinc-binding pocket are highly conservative, but Cys27 and Cys30 are not. The corresponding residues of Cys27 and Cys30 in ENSMFAP00000017249.1, ENSCANP0000009254.1, ENSANAP0000004977.1, ENSHGLP00000008109.1, and ENSSBOP00000033444.1 are replaced by other residues partially or completely, which may result in weak binding to the metal ion. On the other hand, the side chains of Thr32, Thr35, and His36 and the backbones of Tyr33 and Arg49 forms hydrogen bonds with RNA. The mutation of Tyr33 and Arg49 may not reduce the binding affinity, but the change of residues at the position of Thr32, Thr35, and His36 may influence it, such as observed in ENSMFAP00000017249.1, ENSCANP0000009254.1, ENSMICP00000032926.1, and ENSOGAP00000022219.1. For these genes, they all have paralogous genes in certain species with a conserved binding domain similar to other U1C genes. All these results reveal that the animal U1C genes are conserved except for some specific genes and the biological function of these “specific genes” is redundant with other genes of the same organism.
Previous studies have suggested the the plant U1C genes us conserved. The conservation of animal and plant U1C were further compared in this study, and the multiple sequence alignment of animal and plant U1C sequences are shown in Figure S3. It seems that the C-terminal residues of plants are less conservative than those of animals. For example, the residues at the position of Asp57 and Glu64 of human are less conservative (Figure S4). Moreover, some residues are species-specific just as Thr44/Q23, Cys46/Asn25, Ser47/Ala26, Arg49/Tyr28, Glu 53/Ala32, Lys56/Arg35, Lys61/Gln40, Trp62/Phe41, Met63/Glu42, Glu65/Gln44, Ala67/Thr46, Lys73/Gln52, Thr74/Arg53, and Thr75/Ile54 in the animal/plant (the residue number of Arabidopsis U1C). Only Thr44/Q23 is on the interface of interaction surface between U1C and RNA. The Q23 of Arabidopsis U1C prefers to interact with RNA with an extra hydrogen bond, which may improve the binding capability. In summary, even though the U1C genes are often characteristic of the species, the binding surface of U1C is relatively conserved.

In order to investigate the functional relationships between proteins, protein-protein interaction networks of U1Cs was performed on webtool STRING website. In this work, three representative U1C protein sequences of human, mouse and Saccharomyces cerevisiae (yeast) were chosen to generate interaction networks based on experimental inferences (Fig. S5). Interestingly, human and mouse shared 11/11 predicted interacting partners, whereas yeast shared 8/11 (namely NAM8, MUD1 and LUC7) protein interactors with human and mouse (Fig. S5 and Table S3). NAM8, MUD1 and LUC7 are all involved in nuclear mRNA splicing and recognition of 5’ splice site. As expected, all predicted protein interactors of human, mouse and yeast U1Cs play important roles in pre-mRNA splicing. However, specific interaction studies and functional verification requires further analysis between U1Cs and their predicted protein partners.

**Analysis of gene structure and cDNA conserved motifs**

In order to investigate the correlation between the genetic structural characteristics and potential function of animal U1C gene family, it is necessary to compare the gene structure and analyze the presence of cDNA conserved motifs. Accordingly, their genomic organization and corresponding predicted conserved motifs were attached to the vertical phylogenetic tree (Fig. S6). In this work, the sequence of each U1C gene with the longest coding sequence (CDS) was selected to display the exon-intron organization (Fig. S6, middle panels). Exons of U1C genes varied in number from one to seven, which suggested a large difference in gene structures of U1Cs. 48 sequences out of 110 U1C family genes have 7 exon-6 intron gene structure layout, accounting for 35.5% of the total number of members; 23 members have 2 exon-1 intron organization while 18 sequences did not contain any intron sequences (Fig. S6, middle panels). Moreover, only 4 U1C genes (ENSRNOP0000000586.6, ENSXETP00000053216.1, ENSONIP00000018579.1, ENSTRUP00000055830.1) has an extra exon which wasn’t a coding exon. Usually, sequences clustered in the same subclade has similar exon-intron structures such as six members from fish (light blue). Furthermore, sequences from one species may have different gene structures, for example, two sequences from Sarcophilus harrisii were found, where one has one exon (ENSSHAP00000003382.1) and the other one contains 7 exons (ENSSHAP0000001057.1), respectively.
On the other hand, the 10 most conserved motifs were identified based on the cDNA sequence of U1Cs using web-server tool for motif analysis, MEME (Fig. S6, right panel). Broadly, more than half (67/110) of the U1C sequences contained ten conserved motifs. Particularly, other species at the basal part of the tree including Bpp0083134 (Drosophila melanogaster), F08B4.7 (Caenorhabditis elegans), YLR298C_mRNA (Saccharomyces cerevisiae) and ENSCINP00000035879.1 (Ciona intestinalis) displayed low conservation in terms of motif composition, suggesting some degree of divergence in functions. Interestingly, it was observed that no correlation was found between gene structures and conserved motifs. For example, sequences of ENSMMUP00000050488.1 (Macaca mulatta) with one exon, ENSRBIP00000024910.1 (Rhinopithecus bieti) with 2 exons, ENSPTRP00000071198.1 (Pan troglodytes) with 3 exons, ENSNOMP00000021117.1 (Dasypus novemcinctus) with 5 exons, ENSOARP00000011602.1 (Ovis aries) with 6 exons and ENSSSCP00000055297.1 (Sus scrofa) with 7 exons all contained the identified ten conserved motifs (Fig. S6 and Table S2).

**Transcript isoforms and conserved splice sites analysis**

In order to study splicing patterns and conserved splice sites of animal U1C family genes, alternative splicing analysis among animal U1C genes was carried out. Finally, a total of 61 transcript isoforms from 26 U1C genes were obtained from the Ensembl database and linked to the phylogenetic relationships among selected species (Fig. 3, left and middle panels). It was observed that 19 U1C genes contained two transcript isoforms, five other contained three isoforms and finally two contained four isoforms. Furthermore, MEME identified conserved protein motifs from potential protein products from above transcript isoforms are illustrated (Fig. 3, right panel). Expectedly, the primary transcript possess the longest peptide sequence and most conserved motifs while other alternative transcripts have shorter protein length and contained reduced number of motifs. Furthermore, it was observed that alternative first exons (AFE) and alternative last exons (ALE) are the prominent AS events for U1Cs. Moreover, other splicing events such as exon skipping were also observed in Oryctolagus cuniculus, Bos Taurus and so on.

Conserved splice sites or conserved sequences were further identified. Four representative splice sites (Fig. S7A) were identified by using 40-bp flanking sequence at exon-intron junctions (Figs. 3 and S7). Specifically, 3’ splice site (marked in blue arrow) and 5’ splice site (marked in red arrow) of exon skipping events displayed high conservation in Oryctolagus cuniculus, Bos Taurus, Bos Taurus and Equus caballus (Figs. S7B and S7C). Furthermore, type 3 and type 4 of conserved splice sites (marked in purple and pink arrows respectively) were found in the placentals including primates, rodents/lagomorphs, and other mammals (Figs. S7D and S7E). In detail, it was found that type 3 is conserved in ‘primates’ (Fig. S7F) and ‘Other Mammals’ (Fig. S7H) while type 4 is conserved in ‘primates’ (Fig. S7I), ‘rodents and lagomorphs’ (Fig. S7J) and ‘Other Mammals’ (Fig. S7K).

**Expression profile of animal U1Cs**

In order to further investigate the expression profile and regulatory mechanisms of animal U1Cs in a variety of biological aspects such as developmental stage, different tissue and cell type and disease
condition, the expression pattern of model organism (human and mouse) \( U1Cs \) were analyzed. In this work, we mainly focused on the expression of \( U1C \) genes in digestive diseases or in the digestive system (Fig. 4). In detail, human \( U1C \) gene was found to be expressed in lung, liver, thyroid gland, stomach, skin and ovary at a relatively high level according to ‘Pan-Cancer Analysis’ (transcriptomics) (Fig. 4A). Moreover, ‘Tissues, developmental stages - Human - liver’ showed the expression of human \( U1C \) gene was highest in infants, followed by adults, and lowest in adolescents (Fig. S11). In mouse, tissue-specific expression profile from multiple datasets showed that \( U1C \) gene maintained low expression level in various digestive organs including intestine, liver, pancreas, spleen and stomach etc. (Fig. 4B). Moreover, it was observed that transcripts of mouse \( U1C \) were expressed highly in common lymphoid, fetal liver and T cell than in the granulocyte, megakaryocyte and natural killer cells. With regards to the developmental stages of mouse, \( U1C \) accumulated preferably at the fetal stage, higher than its expression in other developmental stages.

Furthermore, all the experiments currently available in Expression Atlas were also analyzed. In human, \( U1C \) gene was highly expressed in several breast, rectal and colon cancer datasets (Fig. S8). Proteome of ‘Wang et al. 2019’ showed that human \( U1C \) protein was highly expressed in prostate, brain, fallopian tube, ovary, lymph node and heart (Fig. S9). In mouse, proteomic map of ‘Organism part - Geiger et al. XXX’ showed that \( U1C \) protein was highly expressed in placenta, preoptic area, prostate gland, renal medulla, saliva-secreting gland and skeletal muscle (Fig. S10). It was also observed that mouse \( U1C \) maintained higher expression level in various sampling time points, strains, developmental stage and somite stages than in various cell type (Figs. S12 and S13).

**Discussion**

**Assessment of phylogenetic relationship and functional conservation in animal \( U1Cs \)**

In the present study, we systematically identified 110 \( U1C \) genes from 61 animal species and reconstructed the phylogenetic relationship of these selected genes. Clear topology showed that \( U1C \) proteins can be broadly divided into four groups including other species, fish, marsupials, monotremes and reptiles and placental, which correlates well with the evolution of animal lineage (Figs. 1 and S2 left panel). Moreover, although over 65% (72/110) of these genes contain multiple copies of \( U1C \) gene (Table S1), analysis of protein structures and motifs of cDNAs and protein domains suggested that this gene family maintains conserved functions, indicating their functional redundancy (Figs. S2 and S6). Furthermore, conserved splicing pattern of animal \( U1Cs \) was analyzed (Figs. 3 and S7) and the majority of transcript isoforms of animal \( U1Cs \) tended to form proteoforms with N-terminal truncation (Fig. 3). However, the truncation of N-terminal does not seem to cause the lack or truncation of the main protein domain (zf-U1), suggesting \( U1C \) proteins may preserve conserved function. Interestingly, exon skipping event was found in vertebrates according to conserved alternative splice site analysis (Fig. S7). Previous studies showed that exon skipping was heavily used for the treatment of Duchenne muscular dystrophy (DMD) [27–29] and \( U1 \) snRNP was also reported to link to congenital myasthenic syndrome (CMS) [13],
thus relationship of the conserved AS event and biological function of U1C needs to be further investigated and experimentally validated.

**Functional diversification of animal U1Cs based on their expression profiles**

U1 snRNP-specific U1C protein is a pivotal component for 5’ splice site recognition and during the early spliceosome assembly. Previous studies have demonstrated that U1 snRNP and U1-70K are involved in Alzheimer’s disease [10]. Here, we found that AD (early and late stage) affected tissues such as amygdala, entorhinal cortex, middle frontal gyrus and superior temporal gyrus exhibited slightly elevated levels of U1C protein compared to non-diseased normal brain (Fig. S9) [30]. Thus a prominent research direction is to investigate the link between U1C levels and AD pathology. Furthermore, expression data shows that U1Cs from human and mouse displayed different expression patterns across different cell types, organs and developmental stages (Figs S8-13), indicating the species- and tissue-specific regulation at the transcription and translational level. However, expression profile of U1Cs at isoform level hasn’t been studied in detail till now and this is something which can be further investigated by quantitative real-time PCR or proteomics approach [31, 32]. Moreover, based on proteomics datasets [33, 34], high expression of U1C was observed in breast, colon and rectal cancers, suggesting its potential involvement and functional role in the development of cancer in these organs (Figs. 4 and S8).

**Comparison of U1Cs in animals, yeast and plants**

Interestingly, splicing machinery is not exactly similar among human, yeast and Arabidopsis. In particular, although the genomic structure of human and plant U1Cs were similar (Fig. S14A), different domains/conserved amino acid sequences were observed at their C-terminal regions (Fig. S14B). This suggests that U1C proteins are conserved at the N-terminus and even at the splice sites located between exon1 and exon2 (Fig. S14B, C). Thus, the functional diversification is potentially mediated by its C-terminal regions.

**Conclusion**

In this study, we identified a total of 110 U1C genes from 61 animal species and analyzed comprehensively their phylogenetic relationship, genomic organization, motif & protein domain enrichment and splicing pattern conservation, providing a foundation for molecular research on U1C proteins for their role in human disease investigated in mammalian cell lines or animal models.

**Declarations**

**COMPETING INTERESTS**

The authors have no conflicts of interest to declare.
AUTHOR CONTRIBUTIONS

F.Y.Z., M.X.C. designed the experiments. K.L.Z., J.L.Z., Y.Z.Z, and J.F.Y. performed the experiments. K.L.Z., J.F.Y., D.D., and M.X.C. analysed the data. K.L.Z., and D.D. wrote the manuscript. G.F.H., S.M.Z., C.E.W. and J.H.Z. critically commented on and revised the manuscript. All authors have read and approved the manuscript.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Jiangsu Province (SBK2020042924), National Natural Science Foundation of China (NSFC31701341), NJFU project funding (GXL2018005), the Natural Science Foundation of Hunan Province (2019JJ50263) and Hong Kong Research Grant Council (AoE/M-05/12, AoE/M-403/16, GRF14160516, 14177617, 12100318).

References

6. T. Lehmeier, K. Foulaki, R. Luhrmann, Evidence for three distinct D proteins, which react differentially with anti-Sm autoantibodies, in the cores of the major snRNPs U1, U2, U4/U6 and U5, Nucleic acids research 18 (1990) 6475-6484. 10.1093/nar/18.22.6475.


