UPPSALA
UNIVERSITET

Differential expression and function of *fubl-1* gene isoforms in *C. elegans*

Joel Pålsson

Degree project in biology, Bachelor of science, 2022
Examensarbete i biologi 15 hp till kandidatexamen, 2022
Biology Education Centre and Institution for Cell- and Molecular Biology, Uppsala University
Supervisor: Andrea Hinas

Abstract

Alternative splicing is the process of producing a variety of transcripts from one and the same gene. This adds further possible variability to gene expression and can in theory mean that one protein coding gene can produce multiple proteins with potentially different functions. Therefore, to understand the function of a gene, alternative splicing must be accounted for. However, this is made more complex by the fact that the existence of different messenger RNA isoforms does not necessarily entail different protein isoforms, which in turn means that an analysis of both the transcripts and final protein is necessary. Far Upstream Element Binding Protein 1 Like 1 (FUBL-1, or C12D8.1) is an RNA binding protein in *Caenorhabditis elegans* which is believed to take part in gene regulation, and which seemingly interacts within an argonaut effector pathway called ERGO-1. The gene has five proposed isoforms for which there are varying amounts of RNA data but only the first isoform, FUBL-1a has proteomics data available. In other words, different messenger RNA isoforms exist but it is unclear which are translated into protein. In this study, I have looked at *fubl-1* and its isoforms to gain further understanding of this protein. This entailed both analysing long read RNA sequencing data to identify messenger RNA isoforms as well as a laboratory analysis of the protein to look for protein isoforms. I found evidence for all isoforms existing as messenger RNAs, and *fubl-1a* was by far the most highly expressed. In my protein analysis, I found indications of different isoforms, but not conclusive evidence.

List of abbreviations

FUBL-1 = Far upstream element binding protein 1 like 1
mRNA = messenger RNA
NLS = Nuclear localization signal/sequence
RNAi = RNA interference
SL = Splice leader
UTR = Untranslated region

Table of Contents

1. Introduction

1.1. Alternative splicing

Gene expression is a complex process spanning from transcription of DNA to the translation of proteins if the gene in question is protein coding. The processes in between these steps are many and complex, many of which are regulated by an equally complex set of molecular interactions (Krebs *et al.* 2018). In eukaryotes, one process which occurs between transcription and translation is splicing. During this, parts of precursor messenger RNA (pre-mRNA) are removed, so called introns, and the remaining merged, called exons. This process in turn is made evermore complex by alternative splicing.

Alternative splicing is the process where one transcript can be altered in ways to produce several different variations of said transcript, so called isoforms (Wang *et al.* 2015). The number of possible isoforms from a gene can range from two to thousands (Wojtowicz *et al.* 2004). However, a gene may have different mRNA isoforms which are not all equally expressed or translated. For instance, when Ezkurdia *et al.* (2015) compared human proteomics data to transcriptomic data, they found that most proteins mapped to only one isoform, suggesting that these have one dominant isoform. This highlights the importance of experimentally providing evidence of isoform existence both as transcripts and proteins.

There are multiple approaches to investigating the isoforms of a specific gene. One method is using RNA sequencing, which will mainly provide transcripts of mature mRNA (i.e. mRNA which has been spliced). Sequencing technologies are commonly divided into long-read and short-read methods, which provide different benefits and disadvantages (Hu *et al.* 2021). Short-read sequencing produces many short sequences which are then computationally reconstructed to give a complete picture. Although possible, an analysis of isoforms based on this risks missing vital information since the reads will rarely span over an entire transcript and precise identification of start and stop sites can be difficult (Conesa *et al.* 2016, Zhao *et al.* 2019). Long-read technologies such as PacBio provides fewer but longer intact sequences, which is beneficial when analyzing mRNA isoforms as it provides a less fragmented view of the transcripts.

1.2. Caenorhabditis elegans

The nematode *Caenorhabditis elegans* is a common model organism within molecular biology and pharmaceutical research. *C. elegans* is easy to work with and cultivate, has a fixed number of cells as adults, and homologs for 60-80% of human genes has been identified (Kaletta & Hengartner 2006). The nematodes have a life cycle of three days (at 25°C), which includes an embryo stage, four larval stages (L1-L4), followed by an adult stage (Corsi *et al.* 2018). *C. elegans* is highly susceptible to RNA interference (in fact, the process was first identified in *C. elegans*), and resources such as Wormbase provide vast amounts of genetic data, genetic models, mutant strains, and other genetic resources freely available for the scientific community (Wormbase. https://wormbase.org).

Like most eukaryotes, *C. elegans* is capable of alternative splicing. Some estimates suggest that 25% of *C. elegans* genes undergo the process, which puts it on the lower end as compared to the 95% alternatively spliced genes in humans. The *C. elegans* genome is similar in intron density to that of vertebrates, but the introns are typically smaller (Gracida *et al.* 2016). Many genes in *C. elegans* also undergo trans-splicing, that is the combination of

two or more exons from different transcripts. In fact, as many as 70% of *C. elegans* mRNAs undergo this process (Blumenthal 2018). The mechanisms of trans-splicing varies and the correct signals are required for transcripts to join together. One type is splice leader (SL) trans-splicing where a small exon is donated to the 5' end of a pre-mRNA by a non-coding RNA. In *C. elegans* there are two types of SL sequences (SL1 and SL2), and these can be a part of operon processing, where it helps stabilize the operon and can separate individual genes, and may aid correct translation by removing upstream start codons (Lasda & Blumenthal 2011).

1.3. FUBL-1

FUBL-1 (Far Upstream Element Binding Protein 1 (FUBP-1) Like 1, C12D8.1) is a protein found in *C. elegans* which is, as the name would suggest, similar to FUBP-1. FUBP-1 is a single-strand RNA and DNA interacting protein in humans (Zhang & Chen 2013). It has been found to act as a regulator of the cell cycle, and is a suspected oncogene (Debaize & Troadec 2019). Previous research has found mutations in FUBP-1 to be more frequent in cells from oligodendrogliomas, i.e. malignant brain tumours, and that FUBP-1 may play a role as oncoprotein in leukaemia (Bettegowda *et al.* 2011, Hoang *et al.* 2019).

In light of this, insights into the function of *C. elegans* FUBL-1 are of vast interest to us. Just as FUBP-1, FUBL-1 has been found to have RNA-binding activity due to the presence of K homology domains (KH-domains), which are known nucleic acid recognition motifs (Kim *et al.* 2005, Valverde *et al.* 2008, Haskell & Zinovyeva 2021). Furthermore, previous studies have shown that FUBL-1 interacts within RNA interference (RNAi) pathways and that the protein interacts with micro RNA (miRNA) targets (Kim *et al.* 2005, Haskell & Zinovyeva 2021).

Additionally, FUBL-1 is believed to interact within the ERGO-1 pathway (Andrea Hinas, personal communication). ERGO-1 is an Argonaut effector (Ago), a family of proteins which bind to miRNAs, Piwi-interacting RNAs (piRNAs), and small interfering RNAs (siRNAs) (Billi 2014). The actions of these RNA-protein complexes vary, but many act through RNA interference pathways (RNAi) by essentially using the bound RNA sequences as guidance to find mRNAs upon which the effector protein can act (Billi 2014). In previous studies, ERGO-1 has been found to target gene duplicates and is thought to prevent overexpression of its targets (Vasale *et al.* 2010, Billi 2014). FUBL-1 deletion mutants have been shown to upregulate the expression ERGO-1 target genes (Andrea Hinas, unpublished observations). This combined with the aforementioned RNA-binding activities suggests that FUBL-1 may have a role in gene regulation in *C. elegans*.

The *fubl-1* gene has five suggested isoforms (Wormbase. https://wormbase.org) (Figure 1), of which only *fubl-1a* has proteomics data available. *ubl-1b* has two proposed transcripts (b1 and b2) which both would translate to the same protein, and *fubl-1c* has been identified via RNA sequencing. The final proposed isoform, *fubl-1d,* is similar to *fubl-1a* but lacks the second exon. This causes a frame shift which results in a short peptide sequence, and the isoform is therefore annotated as non-coding.

Figure 1: *fubl*-1 (C12D8.1) isoforms, 5' to 3'.
Exons are shown in blue, untranslated regions in purple, and introns as grey lines. Grey bars on *fubl-1d* shows a non-coding transcript. Courtesy of Wormbase (Wormbase. https://wormbase.org, accessed 24-04-2022).

Interestingly, a putative bipartite nuclear localization signal (NLS) has been predicted to exist in FUBL-1a. This means that the isoform likely exists both in the nucleus and cytoplasm (Andrea Hinas, unpublished observations). The other isoforms, if translated, lacks this NLS and would not be present in the nucleus. Thus, the presence and function of the protein within a cell would differ depending on the translated isoform.

In addition to this, RNA data on Wormbase suggests the presence of a splice leader sequence (SL1) at the start of the *fubl-1* transcripts (Wormbase. https://wormbase.org). Because this sequence would be present at the start of *fubl-1c*, but within coding or non-coding regions of the other isoforms, this isoform is of certain interest.

1.4. Aim of the study

In this study, I will investigate if there is evidence for the existence of different isoforms of *fubl-1* and if so, in which manner they differ. This analysis will be divided into two:

1. A bioinformatic analysis of RNA-sequencing data to look for mRNA isoforms.
2. A laboratory analysis of FUBL-1 to look for protein isoforms.

2. Material and methods

2.1. *C. elegans* maintenance and strains

C. elegans worms were kept at 20°C and fed *Escherichia coli* OP50 (Brenner 1974). The strains used were N2 wild type (Brenner 1974) and AHS205 with a C12D8.1::3X FLAG, AHS158 FUBL-1 deletion mutant, and AHS170 with a FUBL-1a nonsense mutation. For the protein extraction worms were grown on 2XYT media (1 L: 16 g tryptone; 10 g yeast extract; 5 g NaCl) and fed *E. coli* Na22.

2.2. Bioinformatic analysis

RNA sequencing data produced by Legnini *et al.* (2019. Accession: GSE126465) was obtained from NCBI Gene Expression Omnibus (GEO). The data was downloaded on 01-04-2022 using SRAtoolkit 3.0.0 (SRA Toolkit Development Team. 2022) and aligned using STAR 2.7.9a (Dobin A. 2021). Alignment was done against *C. elegans* genome downloaded from NCBI (accession: PRJNA13758). The alignments were then indexed using SAMtools 1.15 (Genome Research Limited. 2022). Finally, data was analysed using Interactive Genomics Viewer (Broad Institute and the Regents of the University of California. 2022).

After alignment, reads corresponding to *fubl-1* were extracted using SAMtools 1.15 and converted to fasta-files. These were then sorted by strand and analysed for the SL1 sequence (5'-GGUUUAAUUACCCAAGUUUGAG-3') using custom code.

2.3. *C. elegans* collection

C. elegans worms were collected by pouring 1x M9 buffer (22 mM KH_2PO_4; 42 mM Na_2HPO_4; 86 mM NaCl) onto plates, swirling, and then transferred to a tube. The procedure was repeated twice, and the collected worms were centrifuged at 2,800 RCF for 2.5 minutes. The pellets were then washed with 1x M9 between two to seven times depending on growing media (2XYT required more washing), centrifuging at 966 RCF in between until the suspension was clear. The supernatant was then removed and the pellet frozen in liquid nitrogen and stored at -80°C.

2.4. Protein extraction

Mixed stage and embryo N2 and AHS205 worms were collected as described above using three plates for each. The pellets were ground to a fine powder without thawing. 400 µl extraction buffer (50 mM KOH pH 7.4; 150 mM KCl; 5 mM MgCl2; 10% glycerol; 0.1% Triton-X 100; 7 mg/ml Halt protease inhibitor, ThermoFisher Scientific) was added, and the mixture thawed on ice. The thawed suspension was transferred to an Eppendorf tube and stored at -20°C.

2.5. SDS-PAGE and Western blot

Extracted protein from AHS205 and N2 worms were analyzed by SDS-PAGE and Western blot. Three parts protein sample (50 µg protein) was mixed with one part 4X Laemmli buffer (20% 1M Tris-HCl pH 6.8; 0.04 % glycerol; 0.08 % SDS; 0.002 % Bromophenol blue; 20% 2-Mercaptoethanol). The mixed buffer and sample were boiled at 95°C for 5 minutes and put on ice while the gel was assembled.

For the SDS-PAGE, Mini-PROTEAN TGX Precast Gels (Bio-Rad) ran for two hours at 100 V. The gel was transferred using Trans-Blot Turbo Transfer Pack (Bio-Rad) for 30 minutes at 25 V, 1 A. The membrane was then blocked using 3% BSA in TBS-T (50 mM Tris-Cl; 150 mM NaCl; 0.1% NaCl; 0.05 % Tween) overnight on a shaker at 4°C.

After blocking, the membrane was washed in TBS-T, and soaked in mouse monoclonal ANTI-FLAG M2 Peroxidase (HRP) antibody (Sigma-Aldrich) diluted in TBS-T (1:10,000 ratio) at room temperature for one hour. The membrane was then washed again with TBS-T, and equal amount Amersham ECL Prime Peroxide Solution and Amersham ECL Prime Luminol Enhancer Solution (Cytiva) was distributed on the membrane and incubated in darkness for 3 minutes. Images were then captured using a ChemiDoc Gel Imaging System (Bio-Rad).

2.6. RNA extraction

The worms were collected as described above. The frozen pellets were grinded with a micro pestle without thawing and mixed with 900 µl TRIzol Reagent (ThermoFisher scientific). The suspension was further grinded until homogenous and vortexed. For a chloroform phase extraction, 200 µl chloroform was added and the samples were vortexed and left to incubate at room temperature for 10 minutes. After incubating, the suspensions were centrifuged at 24,500 RCF for 15 minutes in 4°C, after which the aqueous phase was transferred to a new tube. The separation procedure was repeated once more using 100 µl chloroform. After the separation, an equal volume of isopropanol was added, and the mix was left to incubate at room temperature for 10 minutes. The supernatant was then removed, and the pellet washed in ice cold 75% ethanol twice. This was centrifuged at 24,500 RCF for 5 minutes, and the ethanol removed. The pellet was then left to dry at room temperature for 15 minutes, and then dissolved in sterile water and stored at -4°C.

2.7. DNase treatment

2 µg extracted RNA was DNAse treated with 1X DNase I buffer and 0.1 U/µl RNase free DNase I (ThermoFisher Scientific). The mixture was incubated at 37°C for 30 minutes, after which 1.5 µl 50mM EDTA (ThermoFisher Scientific) was added and the mixture incubated at 65°C for 10 minutes. This was then put on ice for five minutes, and stored at -20°C.

2.8. cDNA synthesis

Each DNase treated RNA sample was mixed with 8.5 µl master mix, all constituents provided by ThermoFisher Scientific (0.006 µg/µl Random Hexamer Primer; 0.04 µg/µl Oligo(dT) Primer; 2.35 µM dNTPs; 47% 5X Reaction buffer; 1.18 U/µl RiboLock RNase Inhibitor; 23.5 U/µl RevertAid Reverse Transcriptase). The cDNA was synthesized in a single PCR cycle (10 minutes at 25°C, 60 minutes at 42°C, 10 minutes at 70°C).

2.9. RT-qPCR

3 µl cDNA was mixed with 7.5 µl 2X SYBR Green I Nucleic Acid Gel Stain (ThermoFisher Scientific). Primers for *eif-3c* were 5'-ACACTTGACGAGCCCACCGAC-3' and 5'-TGCCGCTCGTTCCTTCCTGG-3', and the primer concentration was 0.15 µM. Primers for *E01G5.4* were 5'-CTCAAGAAAGTTTCACAGCAGGCC-3' and 5'-CACTTACACAAAACATTTCTC-3', and the primer concentration was 0.1 µM. Nuclease-free water was then added to volume. Each sample had three biological replicates, and each biological replicate had three technical replicates.

3. Results

3.1. A majority of transcripts match *fubl-1a*, but all isoforms are present

In this study, I have looked at the *fubl-1* gene and its different isoforms, both as mRNA and proteins. To assess whether there was evidence for differential expression of the *fubl-1* isoforms, a suitable set of sequencing data is necessary. It is preferable if the data is based on a long-read technology, since it allows a better comparison of isoforms. Furthermore, the data had to be transcriptomic (i.e., RNA sequencing) because alternative splicing occurs after transcription. These conditions were met by a set of PacBio RNA-sequencing (RNAseq) data produced by Legnini *et al.* (2019), which contains long-read, mRNA sequencing from L4, and gravid adult worms, respectively. For each worm stage, the data represented two replicates.

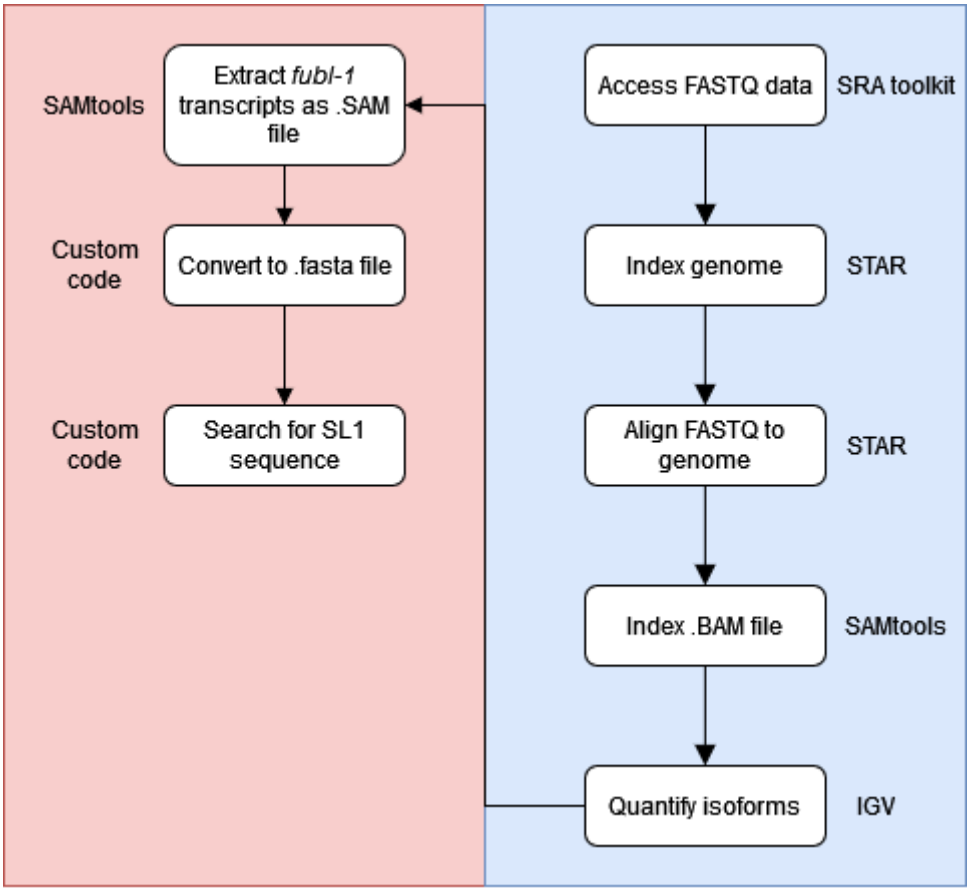

Figure 2: The bioinformatic workflow.
This figure shows the workflow for the bioinformatic analysis, and the tools used in each step. Steps within the blue square shows the process for the isoform analysis, and the steps within the red square shows the process investigating the splice leader sequence (SL1). The custom code used for fasta conversion and SL1 identification was made by Jonas Kjellin.

In order to analyse the data, it had to first be aligned against the *C. elegans* genome, followed by an indexing of the output alignment files (Figure 2). After this, a set of defining features for the different isoforms were identified to quantify them (Figure 3). Both *fubl-1a* and *fubl-1d* could be identified by their respective introns, i.e., the lack of mapped sequences in these regions after alignment. In contrast to these, *fubl-1b1* and *fubl-1b2* have mapped sequences which covers the introns in the other isoforms, and these sequences could be used to define the two isoforms. As seen in Figure 3, *fubl-1c* does not contain any unique feature which could be easily identified bioinformatically and was therefore simply counted manually in Interactive Genomics Viewer (IGV). Because not all reads met one of these criteria, the numbers do not add up to 100%. Reads which did not meet the criteria are mainly sequences which start too far upstream or are too fractured for it to be possible to determine which isoform they match as they lack any of the distinguishing features.

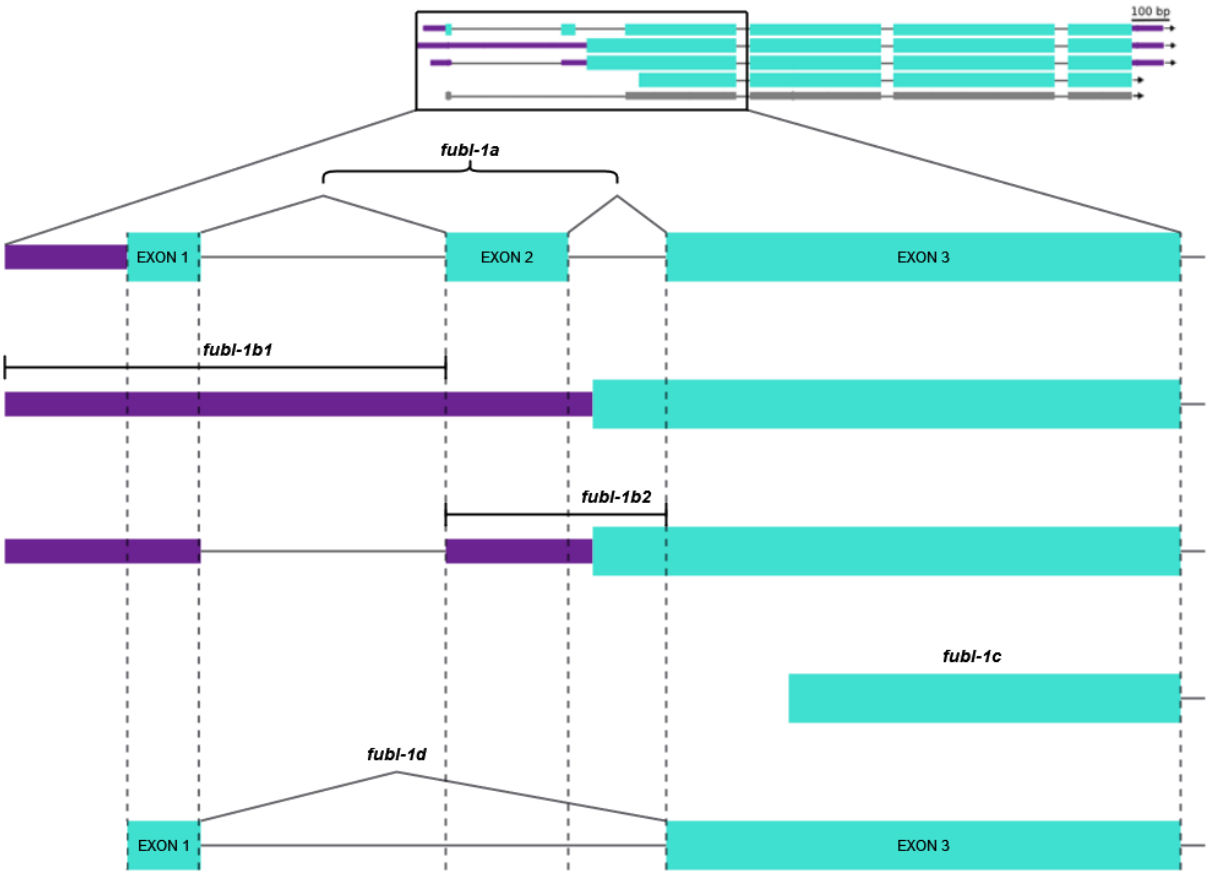

Figure 3: *fubl-1* isoform definition criteria.
Zoomed in view of the first three exons on *fubl-1* highlighting how the different isoforms were identified. Blue bars show coding regions, purple bars show 5'UTR. *fubl-1a* and *fubl-1d* were identified by the presence of their introns: one large for *fubl-1d* and two smaller for *fubl-1a*. *fubl-1b1* and *fubl-1b2* were identified by the presence of sequences in the regions marked by the black lines. Dotted lines highlight the exon positions were they to be present in all isoforms. The image is not to scale.

The two adult replicates had 158 and 127 reads (285 in total), and the two L4 replicates each had 44 (88 in total). In both L4 and adult worms, the data showed overwhelming support for isoform a (Figure 4). For the two adult replicates combined, 81% of the reads match *fubl-1a*, 1% match *fubl-1b1*, 0.7% match *fubl-1b2*. 2.8% match *fubl-1c*, and around 7% match *fubl-1d* (Figure 4B). As for the L4 replicates combined, 68% match *fubl-1a*, 4.5% match *fubl-1b1*, and no reads match *fubl-1b2*. 3.4% match *fubl-1*c and around 14% match *fubl-1d* (Figure 4C).

Almost all reads had a sharp drop in coverage at 10,236,979 bp (chromosome V), as seen in Figure 4. This is most likely a due to a technical issue rather than of biological importance, as such a sequence would not match any suggested isoforms.

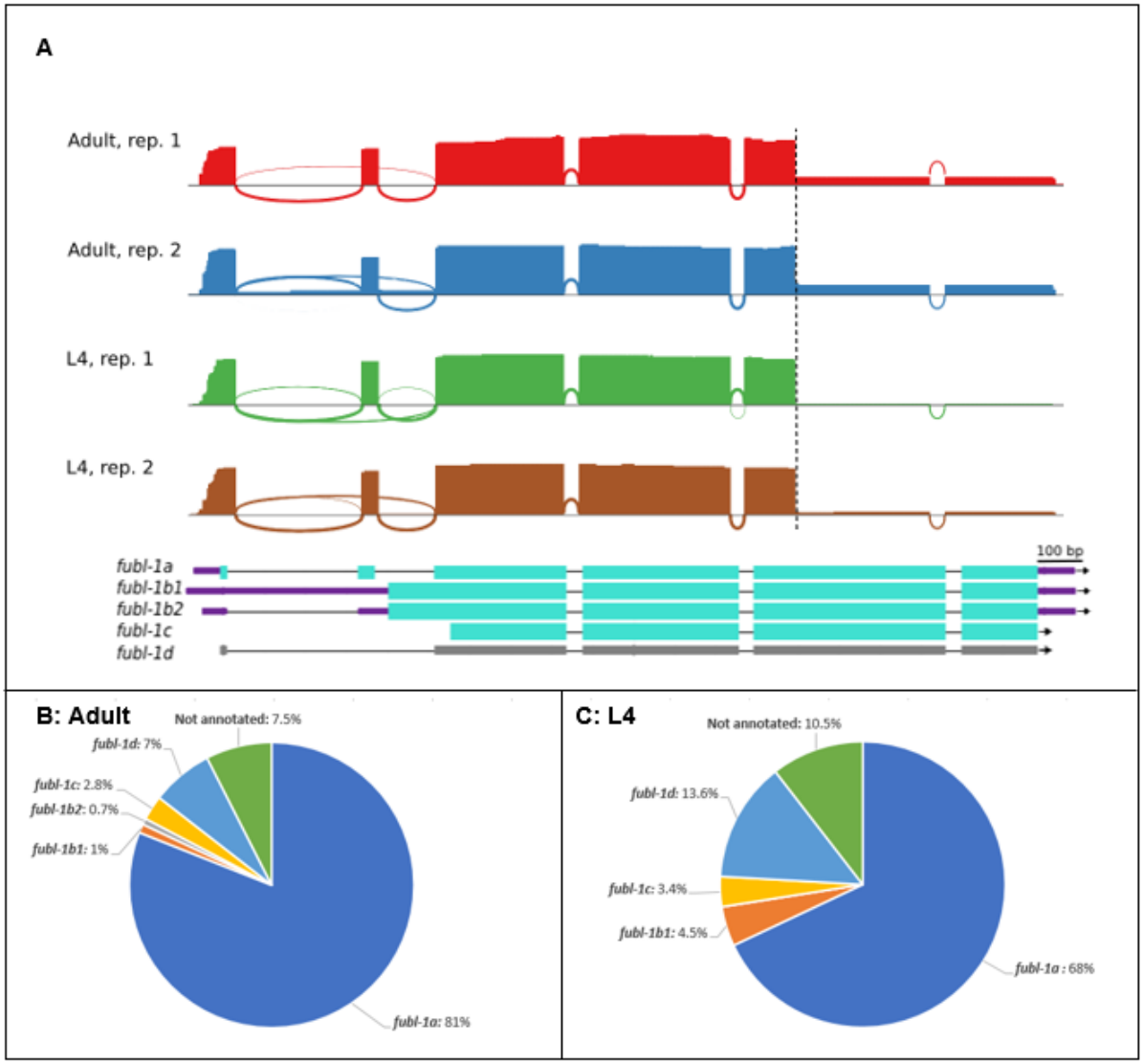

Figure 4: Plot of aligned long-read transcriptomic data as compared to different fubl-1 isoforms, and the distribution of the different isoforms in the RNAseq data.
A: Coverage of RNAseq data from Legnini et al. (2019) as compared to the proposed *fubl-1* isoforms. The height of the bars indicate coverage, and the four different graphs are the four replicates. The dotted line shows where a sharp drop in coverage occurs for all replicates. From the top: adult worms, replicate 1; adult worms, replicate 2; L4 worms, replicate 1; L4 worms, replicate 2. The curved lines show identified introns, and the thickness of the lines indicate the number of reads supporting the intron. Lines on the bottom correspond to the minus strand, and bands on top of it the plus strand. Below the coverage plot are the proposed isoforms of *fubl-1*, courtesy of Wormbase (Wormbase. https://wormbase.org). Accessed 24-04-2022).
B: Percentage of the different isoforms from the adult RNAseq data. **C:** Percentage of the different isoforms from the L4 RNAseq data.

3.2. Alignment of conceptional translations reveal isoform distinctions

To further investigate the isoforms, I compared the conceptual translations to look at the differences in predicted peptide sequences in the first three exons, after which all isoforms are identical (Figure 5). *fubl-1b1* and *fubl-1b2* have different 5'UTRs but have identical peptide sequences, which includes 47 unique amino acids before merging with what is the third exon on *fubl-1a*. *fubl-1a* is the only isoform which includes the second exon as a coding sequence. Both *fubl-1a* and *fubl-1d* start with the first seven amino acids from the first exon, but there is a difference in reading frame which causes *fubl-1d* to encounter a stop codon where other isoforms do not (Figure 5).

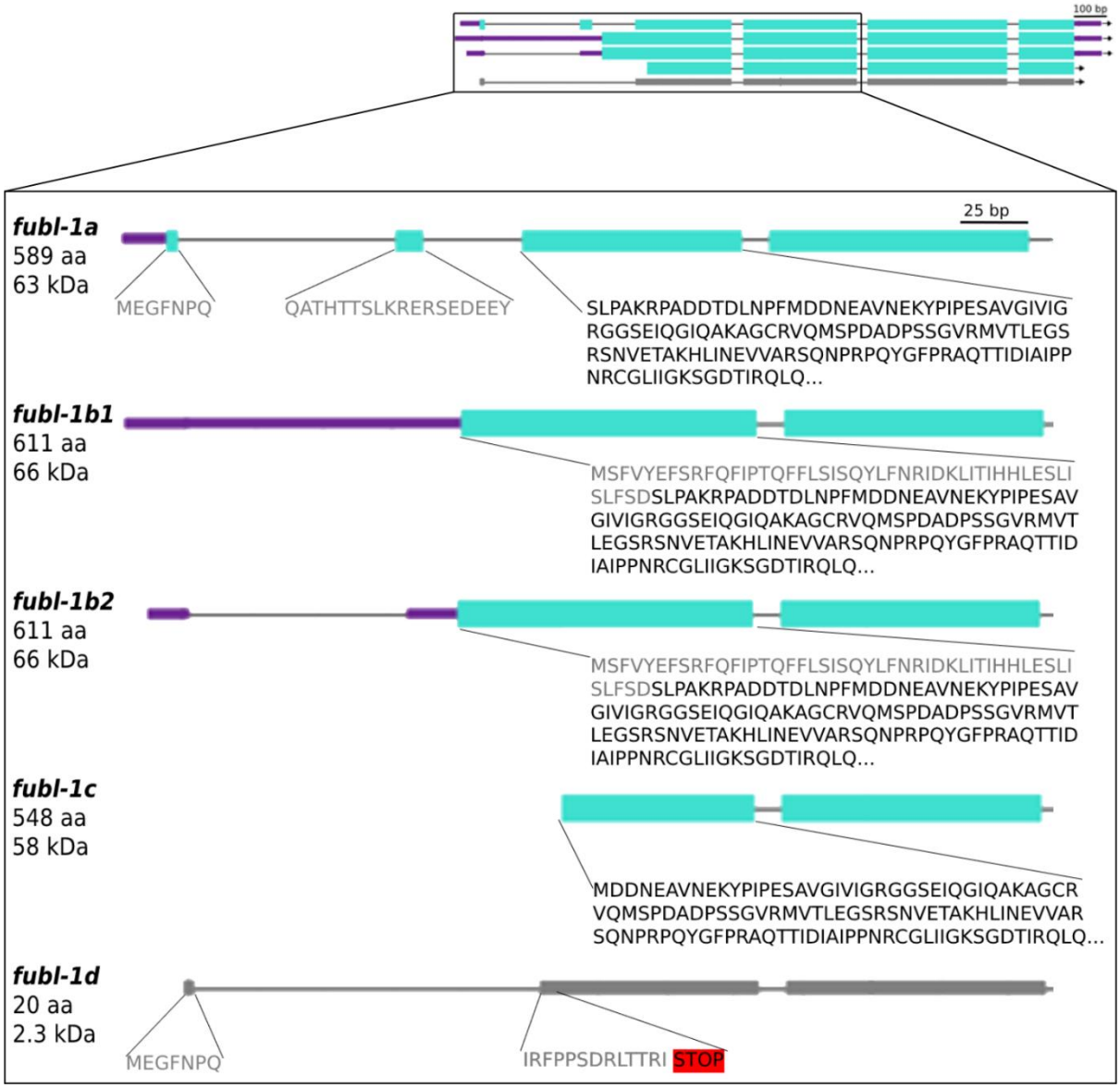

Figure 5: Conceptional translations of proposed FUBL-1 isoforms.
The figure shows the translations of the different isoforms covering the first three exons, as well as the weight and number of amino acids for each protein isoform. Gray text indicates peptide sequences which differ or are not present in all isoforms, black text is common for all isoforms. Exons are shown in blue, 5'UTRs in purple, and non-coding regions in grey.

Finally, to learn more about the structure and function of *fubl-1c*, I looked for the splice leader sequence (SL1) suggested to exist in the 5'UTR region of the isoform (Wormbase. https://wormbase.org). The presence of such a sequence would indicate that *fubl-1c* transcripts are trans-spliced, which would have significant implications on the nature of this

isoform. To do this, reads corresponding to the *fubl-1* gene were extracted, and the SL1 sequence was then searched for within these transcripts (Figure 2), but no such sequence was found. This might be due to sample size, and it is possible that the sequence would be present if more reads were analyzed.

3.3. SDS-PAGE indicate presence of isoforms but provide no conclusive evidence

In order to investigate the isoforms on a protein level, I conducted a Western blot on protein extracted from worms with a FUBL-1::3xFLAG on the C-terminal so as to be present on all isoforms. N2 wildtype was used as negative control so to have something to compare the results with. The predicted sizes of isoforms a-c range between 58 and 66 kDa (Figure 5), which should show up between the 70 and 55 kDa bands on the ruler furthest to the left on the Western. The Western gave four unique bands (Figure 6), one at 70 kDa, another just above 55 kDa, a third just below 55 kDa, and a fourth below 44 kDa. There is also some unspecific binding, appearing as bands outside the expected region on both the control and sample. Taken together, this indicates that FUBL-1 is present in possibly varying sizes, but it is hard to completely distinguish.

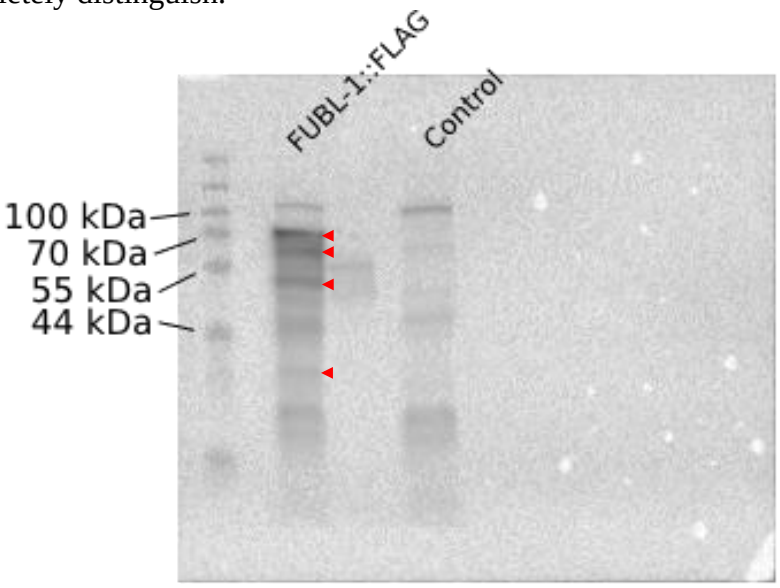

Figure 6: Western blot of FUBL-1 with a FLAG-tag and N2 wildtype showing a difference in signal detection potentially matching that of different FUBL-1 isoforms.
The image is a seven second exposure. The red arrows indicate signals unique to the FUBL-1::3X FLAG. Furthest to the left is PageRuler Prestained Protein Ladder 10 to 180 kDa (ThermoFisher Scientific). The well between the sample and control was not loaded, but there appears to be protein present, which might be carryover from another well.

3.4. Issues with RT-qPCR analysis hinders conclusion

Finally, I investigated to what extent the different isoforms are present as functional proteins. In order to do this, I ran an RT-qPCR with a FUBL-1 deletion mutant (AHS158), a FUBL-1a nonsense mutation (AHS170), and N2 wildtype as control. The target gene, *E01G4.5*, is downregulated by FUBL-1, and by comparing the change in expression of *E01G4.5* between the FUBL-1 deletion mutant and the mutant with a FUBL-1a nonsense mutation, I can estimate to what extent a non-functional FUBL-1a affects the function of FUBL-1. If a disruption of only FUBL-1a affects the target gene expression to the same extent as a complete FUBL-1 deletion, this would indicate that FUBL-1a is the dominant isoform. This requires a reference gene to compare the expression change with, for which *eif-3c* was used.

Unfortunately, the qPCR yielded unreliable data and no conclusion can be drawn (Table 1). The threshold values (Ct) are the number of cycles it takes for the fluorescence to cross the

threshold. These were all higher than expected for both the reference gene and target gene, which indicates low levels of available DNA. There was also high variability between technical replicates, especially for N2 and AHS170 replicates of the target gene. The expression fold change ($2^{-\Delta\Delta Ct}$) is a measurement of the change in target gene expression between the mutants and wildtype, where positive values indicate an increase in expression, and a fold change of 2 equals an increase of 100%. Based on this qPCR, the expression of the target gene barely changes in the FUBL-1a nonsense mutant, which is not in keeping with previous observations, and the Ct data is not reliable enough to be used for any conclusions.

Table 1: Ct and expression fold change data from RT-qPCR.
Each sample had three biological replicates (sample 1-3) and three technical replicates, hence three Ct values for each sample. The reference gene was *eif-3c* and the target *E01G4.5* which is downregulated by FUBL-1. AHS158 is a FUBL-1 deletion mutant and AHS170 a FUBL-1a nonsense mutation. $2^{-\Delta\Delta Ct}$ is the expression fold change. N/A was an undetected value in the qPCR.

	Reference gene				Target gene						
Sample	**Ct 1**	**Ct 2**	**Ct 3**	**Mean Ct**	**Ct 1**	**Ct 2**	**Ct 3**	**Mean Ct**	**ΔCt**	**ΔΔCt**	$2^{-\Delta\Delta Ct}$
N2 1	38.04	37.34	36.32	36.87	N/A	26.98	24.81	33.36	-3.51	6.85	0.01
N2 2	34.97	36.60	36.50	37.08	32.99	25.72	24.70	26.72	-10.36	0.00	1.00
N2 3	37.61	37.32	36.25	36.36	33.73	26.46	24.80	24.80	-11.55	-1.19	2.28
AHS170 1	36.95	35.79	35.86	37.00	30.79	26.30	25.76	30.08	-6.93	3.44	0.09
AHS170 2	36.42	36.12	37.02	36.22	29.89	26.56	25.32	26.43	-9.79	0.58	0.67
AHS170 3	37.06	36.32	36.72	36.87	30.27	24.44	24.78	25.54	-11.33	-0.96	1.95
AHS158 1	35.34	35.38	35.89	35.36	23.70	24.83	25.15	23.72	-11.64	-1.28	2.43
AHS158 2	35.39	35.00	34.47	35.19	23.41	23.67	23.38	23.68	-11.51	-1.14	2.21
AHS158 3	34.62	34.45	35.29	35.59	23.74	23.69	23.59	23.48	-12.10	-1.74	3.34

4. Discussion

In this study I have looked at transcriptomic data of the protein coding *fubl-1* gene in *Caenorhabditis elegans* to see if there is differential expression of its isoforms. In addition to this, I have investigated if there is proof of different isoforms existing as proteins. I find that there are indeed different mRNA isoforms, but it is ambiguous whether these are all translated into proteins.

Based on the RNA sequencing data, it seems as *fubl-1a* is the most common mRNA isoform, but other isoforms are indeed present (Figure 4). This is in keeping with other transcriptional data for FUBL-1 available on Wormbase (Wormbase. https://wormbase.org). It is notable that *fubl-1d*, an isoform which would be translated into merely 20 amino acids (Figure 5), is the second most common transcript in the data by Legnini *et al.* (2019). Its small size is not to say that it lacks biological function however, as short proteins of <100 amino acids have been shown to take part in biological processes (Su *et al.* 2013). Additionally, the isoform could function as a non-coding RNA or perhaps be a way to regulate the expression of *fubl-1*.

I also found transcripts which match *fubl-1c* in the transcriptomic data, but this particular isoform is somewhat difficult to identify as it lacks any unique sequences, introns, or exons. Identifying it by looking for a specific sequence will also match *fubl-1a/b/d* which is why I opted to simply count the isoforms in IGV. This method also has its limitations however, as I risk counting reads from other isoforms which are incomplete or have not been correctly sequenced. Therefore, the exact percentage should be read with caution, but what is certain is that *fubl-1c* does show up in RNAseq data, though the exact frequency is unclear.

Because of the proposed position of a splice leader sequence (SL1) before the start of *fubl*-1c, this isoform was of special interest. It has been found that isoforms which have SL1 trans-splicing show higher translational efficiency than isoforms of the same gene which do not undergo trans-splicing in *C. elegans* by reducing the amount of upstream start codons (Yang *et al.* 2017). However, in my analysis I was unable to find any proof of the SL1 sequence. As of writing (May 2022), all data supporting trans-splicing of *fubl-1* available on Wormbase is based on short-read Illumina sequencing, but no long-read proof exists (Wormbase. https://wormbase.org). This is not to say that the short-read based evidence is incorrect, as the occurrence of the SL1 sequence could simply be too low to show up in the data by Legnini *et al.* (2019). More long-read data might provide proof for the SL1 sequence as well, but in lacking this I can only conclude that if the SL1 sequence is true, its occurrence is rare.

In this study, different life stages were not taken into account; all experiments were either done on mixed stages or on gravid adults. Gene expression often varies between life stages in *C. elegans*, and ERGO-1 is mostly absent in L3, L4, and young adult worms (Spencer *et al.* 2011, Billi 2014). Though the data I analyzed was from L4 and adult worms, I did not conduct a comparative analysis by normalizing the expression towards a reference gene. Doing this might give more insight, and a more systematic analysis of the expression of *fubl-1* throughout the life cycle of *C. elegans* might yield useful data.

If translated, the size of the isoforms would vary slightly from 66 to 58 kDa (Figure 5). The Western blot shows indications that there are differently sized proteins to which the anti-FLAG antibody binds. The first band around 70 kDa does match FUBL-1b. Below is another which can match FUBL-1a, though there is a faint band which matches this for the negative control as well. The band just below 55 kDa is uncertain, as I would expect FUBL-1c to be slightly above instead. Interestingly, a faint band below 44 kDa is seen, and what this corresponds to is unknown. It is unclear how precise this method was, and it is also not clear whether the size difference is large enough to show on a gel of this resolution. More conclusive evidence might be found if a gel of a higher resolution is used, or by doing an analysis based on immunoprecipitation combined with mass spectrometry.

The RT-qPCR did not yield usable data (Table 1). The expected results, in keeping with previous observations, would be a similar increase in expression fold change between the FUBL-1 deletion mutant (AHS158) and the FUBL-1a nonsense mutant (AHS170). The reasons for the unreliable data could be many, as qPCRs have many potential causes of error such as pipetting issues, or an inefficient cDNA synthesis (Taylor *et al.* 2019). Due to availability, only a mutant with a premature stop codon for FUBL-1a was used, but it would be intriguing to redo the analysis with nonsense mutations for the other isoforms as well to see in which manner they may affect the change in expression of the target gene.

4.1. Conclusions

It is likely that FUBL-1a is the most common form of the protein, but there are indications that it might not be the only. As the only isoform with a potential NLS, it being the dominant isoform suggests that FUBL-1 acts within the nucleus. FUBL-1b is similar in structure to FUBL-1a but lacks the NLS and would thus act only within the cytoplasm. Perhaps the function of *fubl-1* is regulated in part by which isoform is translated, which could also explain the relatively high occurrence of *fubl-1d*. It is still unclear whether *fubl-1c* undergoes SL1 trans-splicing. Though all isoforms are present as mRNAs, the ability to draw firm conclusions about the presence of the different isoforms as proteins is halted by the lack of

experimental proof, and further research which provides more insight into this would be valuable.

5. Acknowledgements

I thank my supervisor Andrea Hinas for her help and guidance, and of course for allowing me to do my thesis in her lab. I also extend my gratitude to my fellow student Nouha Abdelaziz for helping me in the lab and kindly providing RNA samples for my RT-qPCR analysis. Finally, I wish to express appreciation to Jonas Kjellin for his invaluable help in my bioinformatic analysis.

6. References

Bettegowda C, Agrawal N, Jiao Y, Sausen M, Wood LD, Hruban RH, Rodriguez FJ, Cahill DP, McLendon R, Riggins G, Velculescu VE, Oba-Shinjo SM, Marie SKN, Vogelstein B, Bigner D, Yan H, Papadopoulos N, Kinzler KW. 2011. Mutations in CIC and FUBP1 Contribute to Human Oligodendroglioma. Science 333: 1453–1455.

Billi AC. 2014. Endogenous RNAi pathways in C. elegans. WormBook 1–49.

Blumenthal T. 2018. Trans-splicing and operons in C. elegans. WormBook

Brenner S. 1974. THE GENETICS OF CAENORHABDITIS ELEGANS. Genetics 77: 71–94.

Broad Institute and the Regents of the University of California. 2022, version 2.12.2. Interactive Genomics Viewer (IGV). URL: https://software.broadinstitute.org/software/igv/home

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17: 13.

Corsi AK, Wightman B, Chalfie M. 2018. A Transparent window into biology: A primer on Caenorhabditis elegans. WormBook

Debaize L, Troadec M-B. 2019. The master regulator FUBP1: its emerging role in normal cell function and malignant development. Cellular and Molecular Life Sciences 76: 259–281.

Dobin, A. 2021, version 2.7.9a. Spliced Transcripts Alignment to a Reference. URL: https://github.com/alexdobin/STAR

Ezkurdia I, Rodriguez JM, Pau ECS, Vázquez J, Valencia A, Tress ML. 2015. Most Highly Expressed Protein-Coding Genes Have a Single Dominant Isoform. Journal of proteome research 14: 1880–1887.

Genome Research Limited. 2022, version 1.15. SAMtools. URL: http://www.htslib.org/

Gracida X, Norris AD, Calarco JA. 2016. Regulation of Tissue-Specific Alternative Splicing: C. elegans as a Model System. I: Yeo GW (red.). RNA Processing: Disease and Genome-wide Probing, s. 229–261. Springer International Publishing, Cham.

Haskell D, Zinovyeva A. 2021. KH domain containing RNA-binding proteins coordinate with microRNAs to regulate *Caenorhabditis elegans* development. G3 Genes|Genomes|Genetics 11: jkab013.

Hoang VT, Verma D, Godavarthy PS, Llavona P, Steiner M, Gerlach K, Michels BE, Bohnenberger H, Wachter A, Oellerich T, Müller-Kuller U, Weissenberger E, Voutsinas JM, Oehler VG, Farin HF, Zörnig M, Krause DS. 2019. The transcriptional regulator FUBP1 influences disease outcome in murine and human myeloid leukemia. Leukemia 33: 1700–1712.

Hu T, Chitnis N, Monos D, Dinh A. 2021. Next-generation sequencing technologies: An overview. Human Immunology 82: 801–811.

Kaletta T, Hengartner MO. 2006. Finding function in novel targets: C. elegans as a model organism. Nature Reviews Drug Discovery 5: 387–399.

Kim JK, Gabel HW, Kamath RS, Tewari M, al et. 2005. Functional Genomic Analysis of RNA Interference in C. elegans. Science 308: 1164–7.

Krebs J, Goldstein E, Kilpatrick S. 2018. Lewin's genes 12. 12th edition. Jones & Bartlett Learning. Burlington, Massachusetts.

Lasda EL, Blumenthal T. 2011. Trans-splicing. WIREs RNA 2: 417–434.

Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N. 2019. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. Nature Methods 16: 879–886.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nature Biotechnology 29: 24–26.

SRA Toolkit Development Team. 2022, version 3.0.0. SRAtoolkit. NCBI. URL: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, Reinke V, Petrella L, Strome S, Von Stetina SE, Katz M, Shaham S, Rätsch G, Miller DM. 2011. A spatial and temporal map of *C. elegans* gene expression. Genome Research 21: 325–341.

Su M, Ling Y, Yu J, Wu J, Xiao J. 2013. Small proteins: untapped area of potential biological importance. Frontiers in Genetics, doi 10.3389/fgene.2013.00286.

Taylor SC, Nadeau K, Abbasi M, Lachance C, Nguyen M, Fenrich J. 2019. The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time. Trends in Biotechnology 37: 761–774.

Valverde R, Edwards L, Regan L. 2008. Structure and function of KH domains: Structure and function of KH domains. FEBS Journal 275: 2712–2726.

Vasale JJ, Gu W, Thivierge C, Batista PJ, Claycomb JM, Youngman EM, Duchaine TF, Mello CC, Conte D. 2010. Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. Proceedings of the National Academy of Sciences of the United States of America 107: 3582–3587.

Wang Y, Liu J, Huang B, Xu Y-M, Li J, Huang L-F, Lin J, Zhang J, Min Q-H, Yang W-M, Wang X-Z. 2015. Mechanism of alternative splicing and its regulation. Biomedical Reports 3: 152–158.

Wojtowicz WM, Flanagan JJ, Millard SS, Zipursky SL, Clemens JC. 2004. Alternative Splicing of Drosophila Dscam Generates Axon Guidance Receptors that Exhibit Isoform-Specific Homophilic Binding. Cell 118: 619–633.

WormBase. Nematode Information Resource. WWW-dokument: https://wormbase.org. Hämtad 2022-05-17

Yang Y-F, Zhang X, Ma X, Zhao T, Sun Q, Huan Q, Wu S, Du Z, Qian W. 2017. *Trans* - splicing enhances translational efficiency in *C. elegans*. Genome Research 27: 1525–1535.

Zhang J, Chen QM. 2013. Far upstream element binding protein 1: a commander of transcription, translation and beyond. Oncogene 32: 2907–2916.

Zhao L, Zhang H, Kohnen MV, Prasad KVSK, Gu L, Reddy ASN. 2019. Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing. Frontiers in Genetics 10: