Identification of putative reader proteins of 5-methylcytosine and its derivatives in Caenorhabditis elegans RNA [version 1; peer review: awaiting peer review]

IC Navarro1,2*, Kin Man Suen1,3*, Dalila Bensaddek4,5, Arun Tanpure6, Angus Lamond4, Shankar Balasubramanian6-8, Eric A Miska1,2,9

1Gurdon Institute, University of Cambridge, Cambridge, CB2 1QN, UK
2Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
3School of Molecular and Cellular Biology, University of Leeds, LC Miall Building, Leeds, LS2 9JT, UK
4Laboratory for Quantitative Proteomics, Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dow Street, Dundee, DD1 5EH, UK
5Bioscience Core Labs, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia
6Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK
7Cancer Research (UK), Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Robinson Way, Cambridge, CB2 0RE, UK
8School of Clinical Medicine, University of Cambridge, Cambridge, CB2 0SP, UK
9Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK

* Equal contributors

Abstract
Background: Methylation of carbon-5 of cytosines (m5C) is a conserved post-transcriptional nucleotide modification of RNA with widespread distribution across organisms. It can be further modified to yield 5-hydroxymethylcytidine (hm5C), 5-formylcytidine (f5C), 2′-O-methyl-5-hydroxymethylcytidine (hm5Cm) and 2′-O-methyl-5-formylcytidine (f5Cm). How m5C, and specially its derivate, contribute to biology mechanistically is poorly understood. We recently showed that m5C is required for Caenorhabditis elegans development and fertility under heat stress. m5C has been shown to participate in mRNA transport and maintain mRNA stability through its recognition by the reader proteins ALYREF and YBX1, respectively. Hence, identifying readers for RNA modifications can enhance our understanding in the biological roles of these modifications.

Methods: To contribute to the understanding of how m5C and its oxidative derivatives mediate their functions, we developed RNA baits bearing modified cytosines in diverse structural contexts to pulldown potential readers in C. elegans. Potential readers were identified using mass spectrometry. The interaction of two of the putative readers with m5C was validated using immunoblotting.

Results: Our mass spectrometry analyses revealed unique binding
proteins for each of the modifications. *In silico* analysis for phenotype enrichments suggested that hm5Cm unique readers are enriched in proteins involved in RNA processing, while readers for m5C, hm5C and f5C are involved in germline processes. We validated our dataset by demonstrating that the nematode ALYREF homologues ALY-1 and ALY-2 preferentially bind m5C *in vitro*. Finally, sequence alignment analysis showed that several of the putative m5C readers contain the conserved RNA recognition motif (RRM), including ALY-1 and ALY-2.

**Conclusions:** The dataset presented here serves as an important scientific resource that will support the discovery of new functions of m5C and its derivatives. Furthermore, we demonstrate that ALY-1 and ALY-2 bind to m5C in *C. elegans*.

**Keywords**
RNA, 5-methylcytosine, Caenorhabditis elegans, epitranscriptomics, readers
Introduction
Nucleoside chemical modifications are a common feature in RNA molecules. More than 160 post-transcriptional modifications have been reported since the discovery of pseudouridine in 1957 (Boccaletto et al., 2018; Cohn, 1960; Davis & Allen, 1957). However, the molecular and physiological functions of most RNA modifications remain unknown.

Discovered in 1958 (Amos & Korn, 1958), the methylation of carbon-5 of cytosines in RNAs (m⁵C) is now known to be a conserved and widely distributed feature in biological systems. m⁵C is catalysed by tRNA aspartic acid MTase 1 (DNMT2) and RNA methyltransferases from the Nop2/Sun domain family (NSUN1-7 in humans), and has been shown to be involved in mRNA stability, ribosome fidelity and translation efficiency (reviewed in García-Vilchez et al., 2019). Previous studies showed that m⁵C is subject to hydroxylation and oxidation by alpha-ketoglutarate-dependent dioxygenase ABH (ALKBH1), forming 5-hydroxymethylcytidine (hm⁵C) and 5-formylcytidine (f⁵C) (Huber et al., 2015; Kawarada et al., 2017). Further processing can occur through 2′-O methylation by FtsJ RNA Methyltransferase Homolog 1 (FTSJ1), yielding 2′-O-methyl-5-hydroxymethylcytidine (hm²⁵C) and 2′-O-methyl-5-formylcytidine (f²⁵C) (Huber et al., 2017; Kawarada et al., 2017; Pais de Barros et al., 1996). While the biogenesis of these chemical marks has been elucidated (Kawarada et al., 2017; Navarro et al., 2020), their functions remain largely unexplored.

The identification of readers, i.e. proteins that specifically, or preferentially, interact with RNA molecules bearing a certain modification, has proven to be an important step towards the understanding of a modification’s biological function. For example, the discovery of N⁰-methyladenosine (m⁰A) readers added new layers of complexity to this pathway, revealing context- and stimuli-dependent functions previously not appreciated (Shi et al., 2019). To date, only two reader proteins have been identified for m⁵C. The Aly/REF export factor (ALYREF) was identified in an RNA pulldown experiment and shown to interact with NSUN2-modified CG-rich regions and regions immediately downstream of translation initiation sites. NSUN2-mediated mRNA methylation is interpreted by binding of ALYREF to modulate nuclear-cytoplasm shuttling of transcripts. (Yang et al., 2017). More recently, the Y-box binding protein 1 (YBX1) was shown to recognise m⁵C in mRNAs and enhance its stability through the recruitment of ELAV1 in human urothelial carcinoma cells (Chen et al., 2019). In zebrafish, YBX1 was shown to stabilise m⁵C-modified transcripts via Pabpc1a during the maternal-to-zygotic transition (Yang et al., 2019).

Recently, we used Caenorhabditis elegans as a model to engineer the first organism completely devoid of m⁵C in RNA. We mapped m⁵C with single nucleotide resolution in different RNA species and found that m⁵C is required for physiological adaptation and translation efficiency at high temperatures (Navarro et al., 2020). Here, to further expand our knowledge on the m⁵C pathway in this organism, we produced the first list of putative readers of m⁵C and its metabolic derivatives hm⁵C, hm²⁵C and f⁵C. The putative readers are enriched in proteins with roles in germline development and RNA surveillance. We show that the C. elegans ALYREF homologues ALY-1 and ALY-2 preferentially bind m⁵C in vitro and identify other proteins that contain a conserved RNA recognition motif potentially involved in m⁵C binding.

Methods
Genetics
C. elegans strains were grown and maintained on NGM plates (3 g/L NaCl, 17 g/L agar, 2.5 g/L peptone, 1mM CaCl₂, 5 µg/mL cholesterol, 1mM MgSO₄, 25 mM KPO₄) seeded with Escherichia coli and transferred to fresh plates regularly to avoid starvation (Brenner, 1974). The strains were kept at 20°C, unless otherwise indicated. HB101 strain E. coli was grown in B broth (10 g/L Bactro-tryptone, 5 g/L NaCl) and used as food source (Caenorhabditis Genetics Center, University of Minnesota, Twin Cities, MN, USA). Bristol N2 was used as the wild type strain.

Oligonucleotide synthesis
The synthesis of 2′-O-methyl-5-hydroxymethylcytidine (hm²⁵C), 5-hydroxymethylcytidine (hm⁵C) and 5-formylcytidine (f⁵C) phosphoramidite monomers was performed as described in Tanpure & Balasubramanian, 2017.

RNA oligonucleotides (5’-GCXUCCGAUGXUACCGAG-GCUGAXC-biotin-3’, where X = C, m⁵C, f⁵C or hm²⁵C) were synthesised in collaboration with ATDBio, Southampton, UK. Purity and integrity of all modified RNA oligonucleotides were confirmed by LC-MS analysis.

Protein extraction
Nematodes were grown until gravid adults in 140 mm NGM plates seeded with concentrated HB101 E. coli and then harvested and washed twice in M9 buffer. A final wash was performed in lysis buffer (20 mM HEPES pH 7.5, 150 mM KCl, 1.5 mM MgCl₂, 0.1% IGEPAL, 0.5 mM DTT) and the animals were resuspended in 4 ml of lysis buffer before the addition of a protease inhibitor cocktail (SIGMAFAST protease inhibitor tablets, Sigma). The animals were pelleted by centrifugation at 2,000 rpm for 2 min (Eppendorf 5810) and the supernatant was removed. Samples were frozen in liquid nitrogen drop-wise, using a Pasteur pipette. Frozen droplets were grinded to powder in metallic capsules for 25 sec in a mixer (Retsch MM 400 Mixer Mill). The powder was stored at -80°C until required. The powder was defrosted at 4°C and the lysate was sonicated for 10 cycles of 20 sec, with breaks of 20 sec in between, in a Bioruptor Pico (Diagenode). The sample was centrifuged in an Eppendorf 5430R microcentrifuge at maximum speed for 15 min at 4°C and the protein concentration of the supernatant was determined using a Bio-Rad Bradford protein assay (Bio-Rad 5000006) according to the manufacturer’s instructions.

RNA pulldown
RNA pulldown protocol was adapted from Dominissini et al., 2012. Magnetic beads (Dynabeads MyOne Streptavidin C1, Invitrogen) were washed twice in 100 mM Tris-HCl pH 7.5,
10 mM EDTA, 1 M NaCl, and 0.1% Tween-20, followed by two washes in lysis buffer. In order to reduce non-specific protein binding to the RNA bait matrix, lysates were pre-cleared by incubation with washed beads for 1 h at 4°C with constant rotation prior to the pulldown. Finally, the following pulldown mix was prepared: 2 mg of protein lysate, 50 U of RNase inhibitor Superase-IN (Invitrogen), 500 pmol of 25-mer biotinylated RNA bait, up to a final volume of 250 µl in lysis buffer. The mixture was incubated with constant rotation for 2 h at 4°C and then added to 20 µl of streptavidin-conjugated beads for pulldown for another 2 h at 4°C. The beads were washed 3–5 times in 1 ml of lysis buffer and the bound proteins were eluted for mass spectrometry in 50 µl 4% SDS 100 mM TEAB at 95°C for 10 min.

Sample preparation for protein mass spectrometry
Samples were processed and analysed as described in Bensaddek et al., 2016. Briefly, samples were reduced using 25 mM tris-carboxyethylphosphine TCEP for 30 min at room temperature, then alkylated in the dark for 30 min using 50 mM iodoacetamide. Protein concentration was quantified using the EZQ assay (Thermo Fisher). The lysates were diluted with 100 mM triethyl ammonium bicarbonate (TEAB) four-fold for the first digestion with endoprotease Lys-C (Fujifilm Wako Chemicals). The lysates were further diluted 2.5-fold before a second digestion with trypsin. Lys-C and trypsin were used (1:50 (w/w) enzyme:substrate), carried out overnight at 37°C. The digestion was terminated with 1% (v/v) trifluoroacetic acid (TFA)/C18 Sep-Pak cartridges (Waters) were sued to desalt the peptides. Desalted peptides were dried and dissolved in 5% formic acid (FA).

Reverse-phase liquid chromatography-MS
RP-LC was performed using a Dionex RS LC nano HPLC (Thermo Fisher Scientific). Peptides were injected onto a 75 µm x 2 cm PepMap-C18 pre-column and resolved on a 75 µm x 50 cm RP-C18 EASY-Spray temperature controlled integrated column-emitter (Thermo Fisher Scientific). The mobile phases were: 2% ACN incorporating 0.1% FA (Solvent A) and 80% ACN incorporating 0.1% FA (Solvent B). A gradient from 5% to 35% solvent B was used over 4 h with a constant flow of 200 nL/min. The spray was initiated by applying 2.5 kV to the EASY-Spray emitter and the data were acquired on a Q-Exactive Orbitrap (Thermo Fisher Scientific) under the control of Xcalibur Software version 2.8-280502/2.8.1.2806 (RRID:SCR_014593) in a data dependent mode selecting the 15 most intense ions for HCD-MS/MS. Raw MS data was processed using MaxQuant (version 1.3.0.5) (RRID:SCR_014485) (Cox & Mann, 2008).

Mass spectrometry data analysis
Putative readers were ranked based on the median of the intensity-based absolute quantification (iBAQ) values (Schwanhäusser et al., 2011). Briefly, for normalisation, iBAQ values were divided by the total sum of intensity of each sample and the median of normalised biological replicates was calculated. A value of 1x10^9 (a number 100 times smaller than the smallest value in the dataset) was then added to all median values to calculate fold changes.
Identification of RNA modification candidate readers. (a) Molecular baits were designed bearing modifications in three structural contexts: single-strand, double-strand, and loop. m$^5$C, hm$^5$C, hm$^5$Cm and f$^5$C monomers were synthesised and incorporated into oligonucleotides tethered to TEG-biotin. Blue circles indicate modified positions. Pre-cleared protein lysate from synchronised adult populations was incubated with unmodified or modified biotinilated RNA baits. Streptavidin-conjugated magnetic beads were used to capture proteins that interact with the RNA. Following washes and elution, proteins were processed and identified by RP LC-MS. (b) Scatter plot of proteins bound to modified versus unmodified RNA baits. Median of the normalised iBAQ values were plotted. Proteins uniquely enriched in m$^5$C, hm$^5$C, hm$^5$Cm and f$^5$C pulldowns are highlighted in mustard. Proteins were considered candidate readers whenever the fold-change of iBAQ values over the “beads only” and “unmodified C” controls was >2.0 and >1.5, respectively. n = 3 independent biological replicates.
considered candidate readers if they preferentially bound to the modified-RNA bait, i.e., at least 2.0- and 1.5-fold increased as compared with the beads and unmodified C controls, respectively (Figure 1B).

To narrow down the number of candidate readers, we focused on unique binders, i.e. proteins that bound specifically to a single modified RNA bait. Using these criteria, we found 143, 62, 81 and 33 putative unique readers for mC, hmC, hmCm and fC, respectively (Figure 2A). We used the WormBase Enrichment Suite to perform phenotype enrichment analysis on the lists of unique candidates (Angeles-Albores et al., 2016). This tool analyses a list of genes according to phenotypes that have been reported by researchers to WormBase upon knockout or RNAi, thus suggesting a potential involvement of such genes in specific biological processes. mC unique binders are the most abundant proteins when compared to its oxidised derivatives (Huber et al., 2015; Huber et al., 2017), likely due to higher levels of this modification in RNA. This subset is enriched mostly for germline development phenotypes, and a few RNA surveillance processes. hmCm-unique binders show specific enrichment for RNA processing phenotypes, whereas hmC and fC show exclusively germline-related phenotypes (Figure 2B).

To validate our dataset, we verified whether ALYREF homologues could be found among our ranked mC reader candidates. Yang et al. identified amino acid residues on ALYREF required for binding to mC (Yang et al., 2019). Protein alignments show that some of these residues are conserved in ALY-1 and ALY-2 (Figure 3A). Indeed, ALY-1 and ALY-2 are 1.8-fold enriched across all triplicates of mC pulldowns. We used transgenic strains expressing either ALY-1 (OP502), or ALY-2 (OP217), tagged at the C-terminus with TY1::eGFP::3xFLAG, to confirm these findings (Sarov et al., 2012). We performed pulldown experiments under the same conditions as before and confirmed using western blotting that both ALY-1 and ALY-2 preferentially bind mC in vitro (Figure 3B; Suen, 2022).

Some mA readers use common RNA binding domains to identify modified RNA (Shi et al., 2019). Amino acid residues shown to be important for mC binding in ALYREF also reside in the RNA recognition motif (RMM) domain, which is a known RNA-binding domain (Yang et al., 2017). The C. elegans orthologues ALY-1 and -2 are predicted to have a similar protein fold in the same region (Figure 3C).

Having shown that ALY-1 and ALY-2 preferentially bind mC-modified RNA, we wondered if other RRM-containing proteins were also identified as mC unique binders. We found that three other proteins with one or more RRMs, RNP-4, SQD-1 and PAB-2, bind to mC baits preferentially. RRM domains consist of a four-stranded antiparallel b-sheet with two a-helices packed against it (Maris et al., 2005). Sequence alignments of the RRM domains in ALY-1, ALY-2, RNP-4, SQD-1, PAB-2 and ALYREF show that a high degree of conservation exists across these proteins in the two ribonucleoprotein (RNP) consensus sequences that are important for nucleic acid binding. Interestingly, a phenylalanine residue outside of the two RNPs is highly conserved in all six proteins (Figure 3D).

In summary, we have identified putative readers for mC, hmC, hmCm and fC in C. elegans using proteomics. To verify the relevance of this proteomic dataset, we confirmed that ALY-1 and ALY-2 interact with mC orthogonally. Finally, we identified three candidate proteins that share a conserved RRM domain with ALY-1/2 (RNP-4, SQD-1 and PAB-2), highlighting these as potential bona fide mC readers.

**Discussion**

This work presents a list of candidate readers of the mC oxidative derivatives hmC, hmCm and fC. Stable isotope labelling experiments suggest that hmCm is a stable modification (Huber et al., 2017). In addition, it has been shown that the ratios between hmC, hmCm and fC in the tRNA Leu-CAA pool varies between organisms and tissues, suggesting a dynamic regulation of these marks in different physiological contexts (Kawarada et al., 2017). This could suggest that these modifications carry out cellular functions themselves, rather than being metabolic by-products of a demethylation pathway. Considering the reported function of fC in mitochondrial protein synthesis (Takemoto et al., 2009), it is expected that the oxidation of mC in tRNAs might affect their role in cytoplasmic translation. The identification of readers provides a first step to elucidate these functions. Importantly, pathogenic mutations have been mapped to NSUN2, NSUN3 and FTSJ1, suggesting that imbalances in these modifications might contribute to the pathogenesis of human diseases (Abbasi-Moheb et al., 2012; Freude et al., 2004; Guy et al., 2015; Khan et al., 2012; Komara et al., 2015; Ramser et al., 2004; Takano et al., 2008; Van Haute et al., 2016).

The use of synthetic modified RNAs as a bait for identification of proteins that are either specifically, or preferentially, interacting with ribonucleoside modifications has been applied with success in the past (Chen et al., 2019; Dominissini et al., 2012; Yang et al., 2017; Yang et al., 2019). Rather than focusing on sequence context, we developed baits bearing modifications in diverse structural contexts. In this work, we identified two proteins showing higher binding affinity to modified RNA: ALY-1 and ALY-2.

The genome of C. elegans encodes three members of the Aly/REF family. It has been shown that simultaneous downregulation of these three genes by RNAi does not compromise either viability, or development. In addition, no defects in mRNA export were observed upon simultaneous knockdown of ALY-1, ALY-2 and ALY-3, suggesting that these proteins mediate alternative processes in the nematode, or that their role in transcript export is redundant with other genes (Longman et al., 2003). This is in agreement with reports showing that in Drosophila ALY proteins are not
Figure 2. Phenotype enrichment for proteins identified as unique candidate readers. (a) Venn diagram showing the number of enriched proteins detected by mass spectrometry in pulldowns using RNA with different m$^5$C, hm$^5$C, hm$^5$Cm or f$^5$C. (b) Phenotype enrichment analysis of unique binders of m$^5$C, hm$^5$C, hm$^5$Cm or f$^5$C. Enrichment analysis was performed with the Wormbase Enrichment Suite. Colour code: m$^5$C in orange, hm$^5$C in purple, hm$^5$Cm in red and f$^5$C in blue.

required for general mRNA transport (Gatfield & Izaurralde, 2002). In contrast, ALY-1 and ALY-2 have been implicated in nuclear retention of a specific mRNA. In C. elegans, sex is initially specified by the ratio of X chromosomes to autosomal chromosomes, i.e. XX animals are hermaphrodites and X0 are males. Female fate requires the genes tra-1 and tra-2, while male cell fate requires inhibition of tra-2 activity. Kuersten et al. showed that binding of NXF-2 and ALY-1/2 inhibits nuclear export of tra-2. In support of these findings, RNAi against ALY proteins led to 5–7% female occurrence in the progeny (Kuersten et al., 2004). It remains to be determined whether m$^5$C acts as an intermediate in
Figure 3. ALY-1 and ALY-2 preferentially bind to m^5C-modified RNA in vitro. (a) Protein sequence alignment of ALYREF and ALY-1/ALY-2 using Clustal Omega. Residues in ALYREF required for m^5C-binding previously identified by Yang et al. are highlighted in red box. "**" represents identical amino acids while ":" and "." represent amino acids with strongly and weakly similar properties, respectively. (b) Anti-FLAG western blotting (left) showing enrichment of ALY-1 and ALY-2 following pulldown using m^5C-modified RNA oligos and relative quantification (right). Mean ± SD, quantification performed using ImageJ, n = 2 independent biological replicates. (c) Overlay of ALY-1 (yellow) or ALY-2 (pink) with ALYREF (grey) structures. ALY-1 and -2 structures were predicted using the SWISS-Model platform which were aligned against a ALYREF crystal structure (PDB: 3ULH). The overall RRM folds are conserved in ALY-1 and -2. (d) Sequence alignment of RRM in m^5C unique binders with ALYREF using Muscle. Black squares highlight consensus sequences RNP1 and RNP2 that are found in RRM. "**" represents identical amino acids while ":" and "." represent amino acids with strongly and weakly similar properties, respectively. Amino acids highlighted in pink are identical to the reference sequence (ALYREF).
this process, for example by enhancing binding affinity of ALY proteins to tru-2 mRNA. Notably, our previous report demonstrated that mC is absent or occurs very rarely in the coding transcripts in *C. elegans*, suggesting that ALY-1/2 could be recognising mC in other molecules (Navarro et al., 2020). In addition, the existence of multiple orthologs of ALY genes in *C. elegans* raises the possibility that these proteins could act redundantly or bind differential sets of modified residues in RNA.

### Conclusions

Our study presents the first comprehensive investigation of putative readers of 5-methylcytosine and its derivatives in the discovery of new functions of the RNA mC pathway.

### Data availability

Underlying data


This project contains the following underlying data:
- Data file: proteinGroups.txt
- Data file: iBAQ values
- Supplementary file: Original Western blots

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

### Accession numbers

PRIDE Project: Putative reader proteins of RNA 5-methylcytosine and its derivatives in *Caenorhabditis elegans*. Accession number PXD035761; https://identifiers.org/pride.project.PXD035761

### Acknowledgements

We would like to thank the Gurdon Institute Media Kitchen for their support providing reagents and media.

### References

PubMed Abstract | Publisher Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text

PubMed Abstract | Publisher Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text

PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text