Unravelling the regulatory function of a short 3´UTR sequence in zebrafish and in human cells

Ana Eufrásio[1,3,4,5*], Joana Azevedo[2,3,4,5*], Joana Machado[2,3,4,6*], Alexandre Ferreira[2,3,4,6], Ana Moutinho[2,3,4], Filipe Henriques[2,3,4], Ana Jesus[2,3,4], Joana Tavares[7], Isabel Pereira-Castro[2,3,4], Joana Teixeira[1,3,4], Pedro A. Pinto[2$], José Bessa[1,3,4,#], Alexandra Moreira[2,3,4,8,#]

Corresponding authors: alexandra.moreira@i3s.up.pt and bessa@ibmc.up.pt

* These authors contributed equally for this work.

$ Current address: Institut de Génomique Fonctionnelle de Lyon (IGFL)

[1]Vertebrate Development and Regeneration; [2]Gene Regulation; [3]i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto; [4]IBMC – Instituto de Biologia Molecular e Celular, Universidade do Porto; [5]ICBAS – Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto - Programa Doutoral em Biologia Molecular e Celular; [6]FCUP - Faculdade de Ciências da Universidade do Porto; [7]GenoMed - Diagnósticos de Medicina Molecular SA, Lisboa; [8]ICBAS – Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto

Corresponding authors: alexandra.moreira@i3s.up.pt and bessa@ibmc.up.pt

* These authors contributed equally for this work.

$ Current address: Institut de Génomique Fonctionnelle de Lyon (IGFL)

Abstract

The 3'UTR contains regulatory elements that control mRNA 3' end formation, polyadenylation, mRNA stability and translation, with an important function in gene expression regulation. Upstream sequence elements (USEs) are *cis*-regulatory sequences localized in the 3'UTR upstream of polyA signals (PAS), with a function in mRNA 3' end formation, that have been well characterized in mammals, virus and *Drosophila melanogaster*. Here we studied the function of the 28 nucleotide USE from the *Drosophila's polo* gene in *Danio rerio* and human cell lines. Bioinformatic analysis revealed a subset of zebrafish, mouse and human genes that contain in their 3'UTR a USE associated with a non-canonical PAS. Microinjecting one-cell stage zebrafish embryos with a plasmid containing the USE sequence and its mutated form USEmt downstream of the *Gfp* reporter gene showed that the USE activates *Gfp* expression, revealing a new function for this non-coding sequence in gene expression in vertebrates. Furthermore, microinjection of *USE* RNA reduces the activity of a USE sensor and induced severe malformation of the zebrafish embryo, indicating that the *USE* RNA may act as sponge sequestering RNA binding proteins (RBPs) necessary for USE function and correct mRNA biogenesis during embryogenesis. Additionally, we found that this mechanism is conserved in human cell lines, as overexpression of the USE leads to a decrease in the expression of nine USE-containing endogenous genes. We successfully identified three RBPs that bind to the USE – HuR, hnRNP C and PTBP1 and show that siRNA depletion of PTBP1 causes a deregulation in the expression of some of these genes. Finally, among several SNPs associated to human disease that overlap with the USE consensus, one (rs3087967) is associated with malignant tumor of colon and generates an ectopic consensus of USE in the 3' UTR of POU2AF2/C11orf53 gene, causing a gain of function of the gene's mRNA, suggesting its involvement in colon cancer development. Taken together our results show that the USE sequence is conserved in zebrafish, mouse and human genes, that it controls the expression of human genes by binding specific RBPs and moreover, that SNPs in USE might be involved in the development of human diseases.

Introduction

The 3' untranslated region (3'UTR) of the mRNA is defined by cleavage and polyadenylation of the pre-mRNA 3' end and has a central function in post-transcriptional regulation of gene expression. This level of regulation is achieved by the existence of specific sequences in the 3'UTR that bind RNA binding proteins (RBPs), or that are targeted by miRNAs, controlling the fate of the mRNA (reviewed in [1-6]).

Polyadenylation signals (PAS) are conserved among fish, flies, mammals, and worms [7]. Similarly to mammals, the most common PAS in *Danio rerio* is the canonical AAUAAA, which is present in 62.3% of all genes, and also contains the GU/U-rich downstream sequence [8]. However, about half of pre-mRNAs do not contain a canonical PAS (AAUAAA) [9, 10] and yet, they are efficiently cleaved and polyadenylated. Usage of non-canonical PAS is enhanced by the presence of auxiliary *cis*-elements in the pre-mRNA, such as the upstream sequence elements (USEs). USEs have been well characterized in mammals as pyrimidine or U-rich sequences that modulate cleavage and polyadenylation in an orientation- and position-dependent manner [11-17]. In humans, USEs modulate of PAS selection and 3' end formation efficiency by acting as an extra-platform for the recruitment of *trans*-acting factors, such as RBPs, with an important function in diseases, such as thrombosis [12, 15, 16, 18, 19].

In the 3'UTR of *Drosophila*'s *polo*, two polyadenylation sites originate two distinct mRNAs with different functions [20]. Upstream of *polo*'s proximal PAS there is a highly conserved 28-nucleotides USE, which has a function in increasing Polo protein levels at the kinetochores [12]. Previous bioinformatic analysis has shown that the most conserved region of *polo*'s USE (TTGTTTTT) is present upstream of a PAS in 2.7% of all human genes [12]. Additionally, USEs are more prevalent upstream of weaker PAS than the canonical AATAAA [12], suggesting a regulatory role in enhancing PAS efficiency in fly genes, as it occurs in human genes. Given the biological features that humans and zebrafish share and in the conservation of the polyadenylation mechanism, it can be hypothesized that the

USE sequence is conserved in zebrafish and involved in gene expression regulation. However, USEs have not been identified in zebrafish so far.

Here we identified the first example of a USE in zebrafish and showed its conservation between *Danio rerio* and humans, and its potential implications in human disease. We explored the USE function in zebrafish and, using a GFP reporter assay, we found that USE increases gene expression by acting in a post-transcriptional manner. We demonstrated that microinjection of *USE* RNA in one-cell stage zebrafish embryos leads to the emergence of several developmental abnormalities in the animal. We confirmed that these results occurred since the microinjected *USE* RNA may sequester factors, such as RBPs, crucial for the correct maturation of endogenous mRNAs necessary for the animal development. Moreover, we showed that this mechanism of action is conserved in humans, and we identified HuR/ELAVL1, hnRNPC and PTBP1 as USE-binding RBPs. Collectively, we provide *in vivo* evidence for the function of USE in zebrafish, its conserved mechanism of action in zebrafish and humans as a regulator of gene expression and the potential implications of USE consensus in human disease.

Results

The USE sequence is present in the 3' UTR of *Danio rerio* and *Homo sapiens* genes

We have previously identified an USE in the 3'UTR of the *Drosophila melanogaster polo* that is essential to maintain Polo protein levels required for cell cycle progression [12]. This sequence is a pyrimidine-rich 28-nt sequence, which is present in 5.2% of *D. melanogaster* genes and highly conserved between the 50 million years distant *D. albomicans* [12]. Given its sequence similarity to previously described human USEs, such as *prothrombin F2* USE [12] and its potential role as a regulatory sequence in gene expression regulation, we investigated its prevalence in *Homo sapiens* and in the vertebrate model organisms *Danio rerio* and *Mus musculus*. We thus developed a R-based bioinformatic script [21] to

identify all the genes that contain in their 3'UTR the most conserved region of *polo's* USE (TTGTTTT) located at a maximum distance of 450 nucleotides (nt) upstream of the weak ATTAAA polyadenylation signal, the same genomic structure previously identified in *Drosophila* (Figure 1A). Our bioinformatic search identified the presence of the USE in 2110 human genes, 1648 mouse genes and 714 zebrafish genes, with subgroups sharing orthologues between species (Figure 1B). Analyses of the gene ontology (GO) terms related to biological processes demonstrated that for *Homo sapiens* the USE-containing genes are involved in dynamic processes such as cellular migration, adhesion and communication, and mRNA processing (Figure S1). For *Mus musculus*, USE-containing genes GO terms include regulation of epithelial to mesenchymal transition and cell response to glucose, and for *Danio rerio*, the USE-containing genes are involved in mitotic cell cycle process, protein ubiquitination and several developmental processes (Figure S1). There are 31 USE-containing genes common to the three organisms (human, mouse and zebrafish) (Figure 1B and C) and interestingly, 20 of these genes are deregulated in neurological diseases or cancer. We then asked if there were common GOs that are highly enriched (F.E ≥ 1.5) among human, zebrafish and mouse and there were no common GO terms between the three organisms (Figure 1D). However, we found 10 common GO terms between human and mouse, 4 GO terms between zebrafish and mouse and 2 GO terms between human and zebrafish (Figure 1D-E). These included GO terms such as "positive regulation of cell migration", "mRNA processing", "cell division" (human and mouse), "chromosome organization" (mouse and zebrafish) and "mitotic cell cycle process" and "mitotic cell cycle" (human and zebrafish).

The USE sequence enhances protein expression in a vertebrate model system

Despite the evolutionary conservation between humans and zebrafish, it is still unknown the function of USEs in *Danio rerio*.

To initially investigate the function of the USE sequence in zebrafish gene expression, the constructs pUC19miniTOL-GFP-USE and pUC19miniTOL-GFP-USEmt were generated. These constructs are ubiquitous expression vectors containing the USE sequence, or the USE sequence mutated in 16 nucleotides (USEmt), subcloned downstream of the GFP coding sequence and assembled in a Tol2 transposon (Figure 2A). Microinjection of these constructs in one-cell stage wild-type embryos, simultaneously with *Tol2* mRNA and another transposon that drives expression of mCherry in the muscle (control of transgenesis), showed that, upon 24 hours post-fertilization (24 hpf), embryos microinjected with GFP-USE DNA exhibited more intense GFP fluorescence, while embryos microinjected with GFP-USEmt DNA demonstrated a more diffused and faded fluorescence (Figure 2B). Quantification of GFP intensity (Figure 2D, left) revealed that embryos microinjected with GFP-USE DNA showed a significantly higher expression than when microinjected with GFP-USEmt DNA, indicating that the USE is able to enhance GFP expression. These results are in agreement with previous results in *Drosophila* [12] and indicate that the mode of action of the USE is similar in vertebrates.

To understand if this activating effect took place at a transcriptional or post-transcriptional level, *GFP-USE* and *GFP-USEmt* RNA were *in vitro* transcribed and microinjected in one-cell stage embryos. Despite the fact that the fluorescence in these embryos was far more diffuse than the one observed for embryos microinjected with DNA (Figure 2C), it is still possible to notice and quantify a higher GFP expression in embryos microinjected with *GFP-USE* RNA than with *GFP-USEmt* RNA (Figure 2C and 2D, right), in a similar manner to the results obtained upon the microinjection of GFP-USE DNA (Figure 2B and 2D, left). These results indicate that in zebrafish the USE sequence acts in a post-transcriptional manner.

USE RNA recruits protein factors essential for the correct expression of genes involved in zebrafish development and in human cells function

Transgenic GFP-USE zebrafish lines were generated using the *Tol2* transposon system [22]. Although the expression of GFP is driven by an ubiquitous promoter (CMV), the stable transgenic line exhibits GFP expression mainly in the notochord, muscle, pituitary gland and eye (Figure 3A and 3B), suggesting a possible role of the USE in regulation of gene expression in these specific structures. Speficially, confocal imaging demonstrates the clear, intense and peculiar pattern of GFP expression in the notochord and muscle areas (Figure 3B).

Given the general function of *cis*-auxiliary elements in serving as scaffolding platforms for the recruitment of *trans*-acting factors, namely RNA binding proteins (RBPs) [1-6], it can be hypothesized that microinjected *GFP-USE* RNA sequesters factors that are crucial for the maturation of endogenous mRNAs essential for zebrafish early embryonic development. If the hypothesis is correct, the microinjection of *USE* RNA in the transgenic GFP-USE embryos would lead to a decrease in the GFP expression, in comparison with the microinjection of *USEmt* RNA (USE sequence mutated in 17 nt) (Figure 3C). This expected outcome would occur since the exogenous *USE* RNA would act as a sponge, sequestering RBPs, making them unavailable to bind to endogenous USE-containing genes and, thus, affecting their expression. Microinjection of *USE* RNA was able to decrease GFP expression in comparison with the microinjection of *USEmt* RNA at 5 dpf (Figure 3D to F), particularly in the notochord and in the muscle, therefore corroborating the hypothesis. These results were consistent at 48 hpf, where it was also observed a decrease in GFP expression in the notochord (Figure S2A to S2C). Further observation of the imaging revealed that the microinjection of *USE* RNA led to the development of a patchy GFP expression that occurred significantly less upon microinjection of *USEmt* RNA at 5 dpf (Figure 3D and G) and at 48 hpf (Figure S3A and S3C). Strikingly, 5 dpf zebrafish larvae microinjected with USE RNA exhibited microphthalmia, a condition where either one or both eyes are abnormally small or absent [23]. When the eye size was quantified, a significant

decrease was observed, when compared to larvae microinjected with USEmt RNA (Figure 4A and B). This phenotype was also observed at 48 hpf (Figure S3D and S3E). Apart from these eye phenotypes, a careful evaluation of 24 hpf larvae allowed us to identify other important phenotypes affecting the notochord, heart, head, late gastrulation and tails (Figure 4C-E) that are more prevalent in USE RNA than with USEmt RNA injected embryos. At 5 dpf, malformations in the notochord, head, tail and swim bladder where also detected (Figure S3A, B and C). To better understand the mechanisms that might be causing these phenotypes, we performed immunohistochemistry in embryos microinjected with GFP-USE and GFP-USEmt RNAs by staining with anti-cleaved caspase-3 polyclonal antibody (pAb) and analyzed the presence of cleaved caspase-3 fluorescence signal in embryos with 48 hpf. We observed a more prevalent presence of cleaved caspase-3 staining in embryos microinjected with GFP-USE RNA, when comparing with not injected embryos and embryos injected with GFP-USEmt RNA (Figure 4F and G), suggesting that, in part, these phenotypes are caused by apoptosis. Overall, these results show that addition of exogenous USE RNA causes several developmental abnormalities, possibly by sequestering RBPs that are necessary to bind to the USE mRNA of endogenous USE-containing genes, whose expression is necessary for correct embryo development.

To understand if this mechanism is conserved in human cells we started by asking whether the USE would have an activating effect in *gfp* expression in HeLa cells. Transfection with the pUC19miniTOL-GFP-USE and pUC19miniTOL-GFP-USEmt constructs show that the activator effect is conserved, as there is a ~2-fold increase in GFP expression with the USE-containing construct in comparison to the USEmt construct (Figure 5A).

As all the plasmids contain the strong CMV promoter we reasoned that the USE transcribed from the transfected plasmid may be act as a “sponge”, to which RBPs necessary to bind to the USE in the endogenous USE-containing mRNAs may preferentially bind, resulting in a decrease in gene expression, in a similar manner as the microinjected USE RNA in zebrafish embryos (Figure 3D). Thus, after

transfection, the mRNA levels of nine USE-containing genes selected from the 31 genes (Figure 1C) identified by the bioinformatic script that are common to human, mouse and zebrafish (*KIF20A*, *EXT2*, *CLDN12*, *MBNL2*, *MED18*, *NR5A2*, *CSNK1A1*, *DKC1* and *MAP4K5*) were quantified by RT-qPCR (Figure 5B-J). The results show that the transfection of GFP-USE leads to a significant decrease in the expression of most of these genes in comparison with the transfection of GFP-USEmt (Figure 5B-J). Thus, these results clearly show that this mechanism is conserved in zebrafish and humans.

HuR, PTBP1 and hnRNP C bind specifically to *USE* RNA in human cells

To understand if the *USE* RNA was acting by sequestering RBPs, it was pivotal to identify these players. To investigate this question, initially, a construct containing a fragment of *polo*'s cDNA followed by the weak polyadenylation signal ATTAAA was generated. By PCR mutagenesis the USE sequence was mutated originating the USEmt sequence, where the pyrimidines were substituted by purines (Figure 6A). These constructs were used to perform *in vitro* transcription generating radioactively labeled pre-cleaved pre-mRNAs that were incubated with HeLa cells nuclear extracts to perform UV cross-linking assays. The results show that multiple proteins assemble onto the *USE* pre-mRNA (Figure 6B, lane 1) and that mutation of the USE sequence leads to a decrease in the binding of some proteins (Figure 6B, lane 2).

To determine the binding specificity of these proteins to the RNA, UV cross-linking competition assays were performed, where increasing amounts of unlabeled RNA was added to the mix (Figure 6C). The results showed that using *USE* RNA as competitor (1 pmol, 50 pmol and 150 pmol- Figure 6C, lane 2, 3 and 4, respectively) significantly inhibits the assembly of the ~35 kDa, ~40 kDa and ~60 kDa proteins to the pre-mRNA. This competition is specific, as addition of the same amounts of unlabeled *USEmt* RNA (1 pmol, 50 pmol and 150 pmol - Figure 6C, lane 5, 6 and 7, respectively) did not cause any noticeable changes in the proteins bound to *USE* pre-mRNA. To identify these RBPs, immunoprecipitation assays

after UV cross-linking were performed. As the orthologue of PTBP1, Heph, was found to bind specifically to this USE [12] and given the molecular weights and the previously characterized consensus binding sites we hypothesized that these RBPs could be HuR/ELAVL1, hnRNPC and PTBP1. HuR is a 32 kDa RBP, ubiquitously expressed in all human tissues that recognizes and binds to U-rich and AU-rich elements (AREs) present in the 3'UTR of target mRNAs [24-28]. HnRNP C1/C2 isoforms (41 and 43 kDa, respectively) bind to polyU tracts and assemble on nascent transcripts and modulate several mRNA functions [29-32]. PTBP1 is a 57 kDa protein that acts as a regulator of alternative splicing with a neuronal function [33] and also as polyadenylation regulator [34-36]. Using antibodies against these proteins it was possible to immunoprecipitate the RBPs after UV cross-linking using the *USE* pre-mRNA and HeLa cell nuclear extracts (Figure 6D-F). In contrast, when the UV cross-linking assays were performed with *USEmt* pre-mRNA the proteins were not as efficiently immunoprecipitated (Figure 6D-F) indicating that HuR, hnRNPC and PTBP1 bind to the *USE* RNA in human cells.

In order to map the specific binding sites within the USE RNA for the identified RBPs, specific mutations were introduced (USE-5' mt, USE-3' mt) and the USE sequence was split in three different fragments: USE-(a), USE-(b) and USE-(c) (Figure S4A). These RNA fragments were used to perform *in vitro* transcription and subsequent UV cross-linking assays. Results show that the USE-(a) region allowed the binding of hnRNP C, HuR and PTBP1 (Figure S4B and C). The USE-(b) fragment only enabled the assembly of hnRNP C, while the USE-(c) fragment disrupted the binding of the three RBPs. The sequence GUUUUU present in USE-(a) and USE-(b) is sufficient for the binding of hnRNP C since this protein cross-links with USE-(a) and USE-(b) fragments (Figure S4C), in agreement with previous data describing UUUUU as a binding site for hnRNP C [31, 32, 37]. Altogether, these results demonstrate that hnRNP C, HuR and PTBP1 specifically bind to the USE sequence, suggesting that the interaction between the *trans*-acting factors and the USE might be an important step in the modulation of gene expression.

PTBP1 modulates the expression of USE-containing genes

To investigate the function of PTBP1 in the expression of USE-containing genes, we performed PTBP1 depletion using siRNAs in HeLa cells. A knockdown of approximately 75% was obtained (Figure S5) and the expression of nine USE-containing genes (*KIF20A*, *EXT2*, *CLDN12*, *MBNL2*, *MED18*, *NR5A2*, *CSNK1A1*, *DKC1*, *MAP4K5*) was quantified by RT-qPCR (Figure 7).

We observed that *EXT2, CLDN12*, *MED18* and *DKC1* mRNA levels decreased significantly upon *PTBP1* knockdown while *MBNL2* and *CSNK1A1* mRNA levels increased (Figure 7). These results indicate that PTBP1 is a modulator of the expression of these USE-containing genes.

A SNP in the USE is associated with cancer

In agreement with the mechanism of action of previously described USEs [11, 12, 15] and our observations, the USE sequence act as a platform for the recruitment of *trans*-acting factors, such as RBPs, to regulate gene expression. Thus, modifications in the USE sequence may be involved in the deregulation of gene expression [38] in the context of human disease. To investigate this, we searched for the occurrence of single nucleotide polymorphisms (SNPs) in the most conserved TTGTTTT region of USE-containing genes (Figure 8A). We found 2 genes containing a SNP in the most conserved area of the USE that each of these respective SNPs have been associated to the development of a specific disease (rs3087967; C11orf53/POU2AF2-Malignant tumor of colon [39]) or trait (rs149940960-GREB1-Alanine aminotransferase measurement [40]) (Figure 8B). Importantly, one of these genes, *POU2AF2/C11orf53* has been recently described to act as a coactivator of POU2F3 to maintain chromatin accessibility, therefore having the potential to control genetic programs [41]. Additionally, it has been also shown that depletion of *POU2AF2/C11orf53* using an in vivo xenograft model of small cell lung cancer, repressed tumor growth and delayed progression of disease in mice [41], showing the tumorigenic potential of these gene in specific cellular contexts. The rs3087967 variant is associated to risk of Malignant tumor of colon and reflects the acquisition of an ectopic USE consensus, compared with the non-

risk variant. Therefore, in agreement with our previous results, that show increased gene expression in the presence of USE consensus in the 3'UTR, we asked if the rs3087967 variant could increase gene expression levels. To test this, we cloned the 3`UTR of *POU2AF2/C11orf53*, containing the two variants (risk and non-risk) downstream of GFP, under the control of a CMV promoter. Using the Tol2 system, we performed mosaic transgenesis assays in zebrafish and analyzed the larvae at 4 dpf. Quantifying the GFP levels in gut cells, we observed that the risk variant showed higher levels of GFP expression, when comparing with the non-risk variant (Figure 8C and D). These results suggest that the acquisition of an ectopic USE in the 3`UTR of *POU2AF2/C11orf53* might lead to an ectopic expression of this gene and consequently, contribute to the development of malignant tumor of colon. These results suggest that the USE might have important unidentified roles in the clinical context.

Materials and Methods

Bioinformatic analysis

The NCBI RefSeq transcripts and the respective genes of *Danio rerio* (assembly danRer10), *Mus musculus* (assembly mm10) and *Homo sapiens* (assembly hg19) genomes were retrieved from UCSC browser [42]. The 3'UTR regions were selected and queried for the presence of the most conserved USE sequence (TTGTTTT) followed by pA1 (ATTAAA) spaced at a maximum of 450 bp, using an R-based script. Subsequently, the zebrafish and mouse genes were converted to human orthologs using BioMart [43].

The Venn diagrams were made by using BioVenn. The Gene Ontology terms from human, zebrafish and mouse USE-containing gene lists were obtained using the GO Enrichment Analysis using Panther tools [44]. Only GO terms related to biological processes with a false discovery rate (FDR) <0.05 were considered. The terms were ordered by the fold enrichment (FE) and then by the FDR. The USE sequence data was intercepted by a dataset of SNPs described as linked to human diseases, available at the DisGeNET database [45]. Single-nucleotide

polymorphisms (SNPs) that occurred in the USE sequence and that had a clinical significance were identified.

Oligonucleotides and Antibodies

The list of oligonucleotides and antibodies used in this work are shown in Supplementary Table 1.

Plasmid constructs

pUC19mini-TOL-USEmt constructs contain the USE/USEmt sequences (from Drosophila melanogaster), USE normal variant/USE risk variant (from C11orf53) and the PAS (ATTAAA) located between both Tol2 arms [22]. *GFP-USE* and *GFP-USEmt* pre-mRNAs were synthesized by *in vitro* transcription using pCS2-GFP-USE and pCS2-GFP-USEmt constructs, respectively. pCS2-GFP-USE/pCS2-GFP-USEmt contains a CMV promoter located upstream of the GPF coding sequence and USE/USEmt sequence. In order to microinject GFP-*USE* and GFP-USEmt RNAs in the transgenic GFP-USE zebrafish embryos, *in vitro* transcription using pCS2-USE construct was performed. This construct contains the SP6 promoter upstream of the USE sequence.

pSPT19-*polo* wt, pGEM7-*polo* wt, pGEM7-*polo* wt-DSE, pSPT19-*polo* mt, pGEM7-*polo* mt or the pGEM7-USE1mt-DSE contain the USE/USEmt sequence located downstream of T7 or SP6 promoters, and they were constructed using the pBS-polo and pxb7 plasmids previously described [46], and pGEM-7Zf(+) and pSPT19 as vector backbones.

***In vitro* transcription**

For synthesis of capped mRNA to microinject in zebrafish embryos, the template plasmids were initially linearized with appropriate restriction enzymes. In vitro transcription was performed incubating 1x transcription buffer, 5 mM DTT, 1 mM ATP, 1 mM UTP, 1 mM CTP, 0.5 mM GTP, 2.5 mM G(5')ppp(5')G RNA Cap Analogue (New England Biolabs), 4 µg DNA, 100 U Ribonuclease Inhibitor (NZYtech) and 60U of SP6 RNA polymerase (Thermo Fisher) for 2 h at 37ºC. The DNA was digested by adding 10 U DNase I recombinant (Sigma) and incubating

for one additional hour. Transcripts were purified by performing phenol/chloroform purification.

For UV cross-linking assays, the USE pre-mRNAs were synthesized by performing *in vitro* transcription using pSPT19-*polo* wt, the pGEM7-*polo* wt or the pGEM7-*polo* wt-DSE constructs as templates. The USEmt pre-mRNAs were synthesized by *in vitro* transcription using the pSPT19-*polo* mt, pGEM7-*polo* mt or the pGEM7-USE1mt-DSE. The USE-(a), USE-(b) and USE-(c) pre-mRNAs were synthesized by *in vitro* transcription using constructs with pGEM-7Zf(+) as the parental vector.

In vitro transcription was performed incubating 1 µg of linearized plasmids with 1x transcription buffer (Roche Applied Science), 10 mM DTT, 28.79 U of RNAguardTM Ribonuclease Inhibitor, 0.1 mM GTP, 0.5 mM CTP, 0.05 mM ATP and UTP, 1 mM Cap Analogue (Ambion), 8 µCi of α-32P-[ATP] and α-32P-[UTP] and 20 U of T7 RNA Polymerase or SP6 RNA Polymerase (Roche Applied Sicences). The reaction was incubated for 1 h at 37ºC. The DNA was digested by adding 5 U of DNase I and incubating at 37ºC for 15 min. The transcripts were purified using Illustra ProbeQuantTM G-50 Micro Columns according with the manufacturer's instructions.

UV cross-linking assays

A mixture containing 2 mM EDTA, 1 mM ATP, 20 mM CP, 11.5 U RNAguard^{TM} Ribonuclease Inhibitor, 2.5% (w/v) PVA and 40 µg/mL of tRNA was incubated with 100 cps of radiolabeled RNA and 5 µL of HeLa cell nuclear extracts. Reactions were incubated for 10 min at 30ºC then 1 µL of 2.5 µg/mL tRNA was added. For competition assays 1, 50 and 150 pmoles of unlabeled competitor RNAs were added with labelled RNA. The reaction mixtures were irradiated twice with $96x10^4$ $\mu J/cm^3$ of UV light for 3 min, using the Hoefer UVC 500 Ultraviolet Crosslinker (GE Healthcare Life Sciences). RNase A was added to the samples, for 30 min at 37ºC, to degrade unprotected RNA. Samples were boiled at 95ºC for 5 min in a 2x SDS gel-loading buffer, in order to denature proteins, and separated by electrophoresis in a 10% SDS-PAGE. The gel was fixed a 10% (v/v) acetic acid, 10% (v/v) glycerol

solution for 30 min at room temperature and then dried at 80ºC under vacuum for 2 hours. The radiolabeled proteins bands were visualized by autoradiography.

Immunoprecipitations

For immunoprecipitation of UV crossed-linked proteins with monoclonal antibodies, 400 µL of 10% (v/v) protein A Sepharose CL-4B beads (GE Healthcare Life Sciences) in IP-2 buffer [50 Mm Tris pH 7.9, 50 mM NaCl, 0.1% (v/v) NP-40] was incubated with 40 µg of rabbit anti-mouse antibody (DakoCytomation) in a vertical wheel for 90 min at 4ºC. The beads were washed 3 times with ice-cold IP-2 buffer. One UV cross-linking reaction and 2 µL of anti-hnRNP C/ Elavl1 was added to the beads and the mixture was rotated for 1 h at 4ºC. The beads were subsequently washed 3 times and dried. 20 µL 2x SDS gel-loading buffer was added to the beads and the proteins were denatured at 95ºC for 5 min and separated by gel electrophoresis in a 10% SDS-PAGE. The gel was fixed and dried at 80ºC under vacuum for 2 h and the radiolabeled protein bands were visualized by autoradiography.

Zebrafish husbandry and embryo culture

All zebrafish (*Danio rerio*) experiments were done in accordance with the European animal welfare regulations and standard protocols. Fertilized eggs were kept at 28 ºC in E3 medium supplemented with 0.001% 1-phenyl-2-thiourea to prevent pigmentation development [47].

Microinjection procedures, Zebrafish GFP-USE transgenic line establishment, GFP screening and quantification

One-cell stage AB WT embryos were microinjected with 3nL containing 25ng/ul *Tol2* transposase mRNA and 25ng/ul phenol/chloroform-purified reporter vectors: pUC19-miniTOL-GFP-USE, pUC19-miniTOL-GFP-USEmt and mylz:mCherry, being the latter used as internal control of transgenesis, driving mCherry expression specifically in muscle fibers of the zebrafish. Embryos were screened at 24hpf for GFP expression, using a fluorescence stereomicroscope. The images were acquired using Leica M205 and analyzed using ImageJ software [48].

For the GFP-RNAs microinjection, one-cell stage AB WT embryos were microinjected with 3nL containing 300ng/ul GFP-USE and GFP-USEmt RNAs. Embryos were screened and quantified at 24hpf for GFP expression as described before.

Generation of the GFP-USE zebrafish transgenic line was performed using the *Tol2* transposon system [22]. One-cell stage AB wild-type embryos were microinjected with 3nL containing 25ng/ul *Tol2* transposase mRNA and 25ng/ul phenol/chloroform-purified reporter vector (pUC19-miniTOL-GFP-USE). Embryos were screened at 48hpf for GFP expression, using a fluorescence stereomicroscope, and raised until adulthood. Potential founders were then crossed with WT zebrafish and the progeny with consistent GFP expression was selected and raised until adulthood, establishing this way the GFP-USE transgenic line.

For the USE-RNAs microinjection, one-cell stage AB WT embryos were microinjected with 3nL containing 300ng/ul USE and USEmt RNAs. Embryos were screened and quantified at 24hpf and 5dpf for GFP expression as described before.

The fluorescence was calculated using the following formula [49]: GFP Fluorescence = Integrated Density – (Area of selected area X Mean fluorescence of background readings).

For the assessment of the impact of the rs3087967 risk variant, we performed in vivo mosaic transgenesis assay. One-cell stage AB WT embryos were microinjected with 3nL containing 25ng/ul Tol2 transposase mRNA and 25ng(ul phenol/chloroform-purified reporter vectors (pUC19-miniTOL-GFP normal variant and pUC19-miniTOL-GFF risk variant and mylz:mCherry). Embryos were raised until reach the 4dpf larval stage, fixed and stained with DAPI. The images were acquired using SP5II confocal microscope.

Immunohistochemistry

Transgenic GFP-USE zebrafish embryos were fixed in 4% formaldehyde in PBS1x, overnight at 4ºC, and then washed in PBS-T (0.1% Triton X-100 in PBS1x), permeabilized in PBS-T 0.5%, rewashed and incubated with DAPI (1:1000) (Invitrogen) diluted in PBS-T 0.1%, for 4h. Embryos were extensively washed and mounted in microscopic slides in 50% glycerol in PBS1x. Images were acquired using Leica SP5II confocal microscope and analyzed in ImageJ [48]. USE and USEmt RNAs microinjected embryos were fixed and washed as described before. For permeabilization, the embryos were incubated with 100% cold MeOH and blocked with in situ hybridization buffer. The embryos were then incubated with anti-cleaved caspase 3 pAb (1:100) followed by a wash and a second incubation with anti-rabbit 568. Finally, the embryos were washed, the images acquired in Leica M205.and analyzed in ImageJ [48].

pUC19-miniTOL-GFP normal variant and pUC19-miniTOL-GFF risk variant microinjected embryos were fixed, washed and incubated with DAPI as described before. Images were acquired using Leica SP5II confocal microscope and analyzed in IMARIS software (Imaris (RRID:SCR_007370)).

Cell culture and transfection

HeLa adherent cell line was grown and maintained in Dulbecco's Modified Eagle Medium (DMEM) with GlutaMAX, supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin solution (10000 U/mL) (Gibco, Thermo Fisher Scientific). The cell line was maintained in a humidified incubator at 37º C with 5% CO_2.

HeLa cells were transfected with 500 ng of plasmid DNA using Lipofectamine 2000 (Thermo Fisher Scientific) and following manufacturer's instructions. To knockdown *PTBP1*, two siRNAs targeting this RBP (siPTBP1) were transfected into HeLa cells (siPTBP1_1: 5'-GCACAGUGUUGAAGAUCAU-3'; siPTBP1_2: 5'-AACUUCCAUCAUUCCAGAGAA-3'; Sigma-Aldrich), using the jetPRIME® (Polyplus Transfection) reagent, following the manufacturer's guidelines. A non-target siRNA pool (siNTC: 5'-UGGUUUACAUGUCGACUAA-3'; 5'-UGGUUUACAUGUUGUGUGA-3'; 5'-UGGUUUACAUGUUUUCUGA-3'; 5'-

UGGUUUACAUGUUUUCCUA-3'; Dharmacon) was also transfected, to serve as a negative control. Cells were transfected at 50% confluency with 50 nM of siPTBP1 (25 nM siPTBP1_1 + 25 nM siPTBP1_2) or 50 nM of siNTC. Cells were collected for total RNA extraction 48h post-transfection.

RNA extraction and RT-qPCR

RNA was extracted using TRIzol (Ambion), according to the manufacturer protocol. 1 µg of RNA was treated with Recombinant DNase I (Roche) and cDNA was synthesized using Superscript IV reverse transcriptase (Invitrogen) and random hexamers. RT-qPCR reactions were performed using SYBR Select Master Mix (Applied Biosystems) in the 7500 Fast Real-Time PCR System thermocycler (Applied Biosystems). The list of primers used are listed in Supplementary Table 1 and the results were analyzed using the ΔΔCt method [50] using *18S* as the housekeeping gene.

Statistical Analysis

Statistical analyses were performed using unpaired Student's t-test and by using χ^2 test with Fisher correction. p-values <0.05 were considered as statistically significant. *, p<0.05; **, p<0.01; ***, p<0.001; ****, p<0.0001.

Discussion

USEs are characterized by being *cis*-auxiliary elements capable of modulating mRNAs 3'-end formation [13]. These elements were initially described in virus and mammal genes and, more recently, in *Drosophila melanogaster* [12]. Here, we found that the 28-nt USE described in *Drosophila*'s *polo* gene is conserved in the 3'UTRs of *Danio rerio, Mus musculus* and *Homo sapiens* genes. Previous studies have shown that the deletion of *polo*'s USE sequence led to a decrease in Polo's protein levels *in vivo* [12]. Noticeably, here we showed that microinjection of GFP-USE in one-cell stage zebrafish embryos drives higher GFP expression than microinjection of GFP-USEmt, showing that this USE activate gene expression in

distinct model organisms and demonstrating a functional conservation of this non-coding sequence.

Trans-acting factors bind to *cis*-auxiliary elements to modulate mRNA metabolism [1-6]. We found that exogenous microinjected *USE* RNA acts as a sponge, sequestering factors such as RBPs, making them unavailable to bind to endogenous USE-containing genes, causing several defects in zebrafish development. Therefore, these results suggest that USE-containing genes have a crucial role in the zebrafish development. By transfecting HeLa cells with the same GFP-USE construct we found that the transfected USE acts as a sponge, in a similar manner as in zebrafish, indicating a functional conservation of the USE. Importantly, overexpression of USE in a human cell line leads to modulation in the expression of 9 out of the 31 USE-containing genes common to zebrafish, mouse and humans. We found that *EXT2*, *CLDN12*, *DKC1*, *MED18* and *DKC1* are downregulated by USE overexpression in human cells. Furthermore, siPTBP1 depletion also downregulates these USE-containing genes, revealing the USE-dependent molecular mechanism involved in their expression. Importantly, altered expression of these genes is associated with serious human diseases, particularly in cancer and neurological diseases. *EXT2*, for instance, is a glycosyltransferase deregulated in the EXT2-related syndrome where one of the manifestations of the disease is microcephaly [51]. Notably, this is phenotype we observed in zebrafish embryos microinjected with USE RNA and in human cells *EXT2* is downregulated by overexpression of the USE, revealing a new mechanism of *EXT2* expression dependent on the USE. *CLDN12* (Claudin12) is another USE-containing gene, which codes for an integral membrane protein and is overexpressed in colorectal carcinomas [52]. Our results now clearly show that *CLDN12* can be downregulated by USE overexpression. Remarkably, these genes are downregulated by siPTBP1 depletion, revealing that PTBP1 modulate their expression through binding to the USE. These results reveal the molecular mechanisms involved the expression of genes associated with neurological diseases and cancer suggesting new USE-dependent mechanisms of disease.

The conservation of the function of USE in zebrafish and humans brought us to identify the players involved in the USE regulation. Thus, by performing UV-cross linking assays with human cells we found that HuR and hnRNP C specifically bind to the USE sequence. Additionally, we were able to map the binding sites for both RBPs and we found that HuR binds to the sequence UAUUUGUUUUU and hnRNP C to GUUUUU. Both identified RBPs have crucial roles in mRNA metabolism, namely in splicing, polyadenylation, translation efficiency and mRNA stability [28, 30, 53-55]. Particularly, HuR has a key role in enhancing mRNA stability by binding to its targets [54, 55]. The fact that HuR is able to stabilize transcripts suggests and the increase in gene expression USE-dependent we observed in zebrafish and human cells, it is possible that the molecular mechanism associated with USEs gene expression regulation is increasing mRNA stability.

In this study, we found evidences that the USE gene expression regulation might have an important clinical relevance. We found that the SNP rs3087967, associated to colorectal cancer, is present in the 3`UTR of the gene POU2AF2/C11orf53 and generates an ectopic USE consensus. We also found that this ectopic USE consensus caused by rs3087967 is sufficient to increase gene expression in vivo, in zebrafish gut cells, using a transgenesis reporter assay. Because *POU2AF2/C11orf53* has been shown to have a tumorigenic potential in the context of small cell lung cancer [41], we hypothesize that rs3087967 might lead to an increase of *POU2AF2/C11orf53* expression, contributing to colorectal cancer development. These results highlight the novel role of USE in the mechanisms of human disease.

Overall, our study has unveiled the emerging importance of USE, firstly identified in *Drosophila*, and that now we show to have shown to have a conserved function across zebrafish and humans. We showed in vivo and in vitro that this sequence might have similar regulatory functions across species. Moreover, we were able to identify specific RBPs that bind to the USE RNA and demonstrated their role in human gene expression regulation. Finally, we showed, in vivo, the impact of an

associated disease SNP within the USE, suggesting a novel role in human disease.

Figures Legends

Figure 1- The USE sequence is conserved across human, zebrafish and mouse genes

(A) Workflow used to obtain the Gene Ontology (GO) terms related to biological processes of the common USE-containing genes between the human, mouse and zebrafish transcriptomes obtained from an R-based bioinformatic script, which identified all the genes that contained, in the 3' UTR, the most conserved region of the USE sequence (TTGTTTT) located upstream of the polyadenylation signal (ATTAAA), at a maximum separating distance of 450 base pairs (bp). **(B)** Venn Diagram representing: the number of USE-containing genes found in human (2110), zebrafish (714) and mouse (1648). 80 USE-containing genes are common between human and zebrafish while 31 are common between the 3 species. **(C)** The 31 USE-containing genes that are common between the 3 species: human, mouse and zebrafish. (**D**) Venn diagram representing the common GOs highly enriched between human, zebrafish and mouse (F.E≥1.5). A total of 10 common GO terms were found between human and mouse, 4 between zebrafish and mouse and 2 between human and zebrafish. (E) Table representing the GO terms found to be highly enriched in human, zebrafish and mouse.

Figure 2 - The USE sequence increases GFP expression in a post-transcriptional manner in *zebrafish.*

(A) Schematic representation of pUC19miniTOL-GFP-USE and pUC19miniTOL-GFP-USEmt plasmids used to microinject one-cell stage zebrafish embryos. The black boxes correspond to the cytomegalovirus (CMV) promoter region; the green

boxes correspond to the GFP reporter gene; and the grey boxes represent the USE sequence, or the USE sequence mutated in 16 nucleotides, followed by the polyadenylation signal (PAS)- ATTAAA. **(B)** Representative images from 24hpf zebrafish embryos microinjected with pUC19miniTOL-GFP-USE/pUC19miniTOL-GFP-USEmt with *Tol2* mRNA and mylz:mCherry vector. **(C)** Representative images from 24hpf zebrafish embryos microinjected with *GFP-USE* RNA and *GFP-USEmt* RNA, **(D)**. Representative graph for the GFP expression quantification from microinjected embryos with pUC19miniTOL-GFP-USE and pUC19miniTOL-GFP-USEmt (left); Representative graph for the GFP expression quantification from microinjected embryos with GFP-USE and GFP-USEmt RNAs; Statistical significance was determined by χ^2 test with Fisher correction or by two-tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05. Images were acquired with Leica M205(scale bar = 100 µm) and the GFP expression was quantified for at least 25 embryos in each.

Figure 3 - Microinjection of *USE* RNA causes a decrease in GFP expression in transgenic GFP-USE animals.

(A) Representative image of a 5dpf GFP-USE transgenic line. (**B**) Confocal images of transgenic GFP-USE zebrafish, stained with the nuclear marker DAPI, showing GFP expression in the notochord and muscle.(Scale bar= 10 µm) (**C**) *USE* RNA and *USEmt* RNA sequences used to microinject transgenic GFP-USE zebrafish embryos. (**D**) Representative image of GFP expression in transgenic GFP-USE zebrafish embryos microinjected with *USE* RNA or *USEmt* RNA, at 5 dpf. Images were acquired using Leica M205 (**E**) Quantification of GFP expression in notochord (n=64 USE; n=54 USEmt) and (**F**) in muscle (n=53 USE; n=54 USEmt); **(G)** Quantification of the percentage of transgenic GFP-USE embryos that showed a patchy expression of GFP in the notochord (73%,n=69 USE;50%,n=.59 USEmt) Statistical significance was determined by χ^2 test with Fisher correction or by two-tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05. (Scale bar = 100 µm).

Figure 4 - Microinjection of *USE* RNA in one-cell stage zebrafish embryos induces the development of several defects.

(**A**) Representative graph of the quantification of eye size (pixels) in microinjected embryos with USE and USEmt RNAs, at 5dpf. (n=69 USE; n=58 USEmt) (**B**) Representative graph of que quantification of the percentage of microphthamia in microinjected embryos with USE and USEmt RNAs, at 5dpf (28%, n=68 USE; 7%,n=58 USEmt. (**C**) Quantification of the several defects developed 24h upon the microinjection of *GFP-USE* RNA (in green) and *GFP-USEmt* RNA (in grey) in one-cell stage zebrafish embryos. (**D**) Pie chart representing the total number of defects developed upon the microinjection of *GFP-USE* RNA and *GFP-USEmt* RNA. (**E**) Representative images showing the various abnormalities upon microinjection of *GFP-USE* RNA at 24 hpf, indicated with red arrows. (**F**) Representative images of not injected embryos and in embryos injected with USE and USE RNAs, at 48hpf, showing in purple, the cleaved anti-caspase-3 staining. (**G**) Representative graph showing the percentage of embryos with consistent appearance of apoptotic cells, characterized by the presence of anti-caspase-3 staining (42%,n=39 Not injected; 83%,n=33 USE;59%, n=33 USEmt). Statistical significance was determined by χ^2 test with Fisher correction or by two-tailed unpaired t-test. ***$p<0.001$; **$p<0.01$; *$p<0.05$; ns $p>0.05$ Images acquired with Leica M205.Scale bar = 100 µm

Figure 5 - The USE increases *GFP* mRNA levels in HeLa cells and modulates human USE-containing genes mRNA levels.

(**A**) RT-qPCR showing that the mRNA levels of *GFP* increase by 3-fold upon transfection with the pUC19miniTOL-GFP-USE plasmid (USE) when comparing with the transfection with the pUC19miniTOL-GFP-USEmt plasmid (USE mut). mRNA levels of *KIF20A* (**B**), *EXT2* (**C**), *CLDN12* (**D**), *MBNL2* (**E**), *MED18* (**F**), *NR5A2* (**G**), *CSNK1A1* (**H**), *DKC1* (**I**) and *MAP4K5* (**J**) were assessed by RT-qPCR. Apart from *CSNK1A1* and *MAP4K5*, transfection with the pUC19miniTOL-

GFP-USE plasmid decreases the mRNA levels of the USE-containing human genes. The data is presented as the mean ± SD of at least three independent experiments and analyzed by two-tailed unpaired Student's t-test. ***$p<0.001$; **$p<0.01$; *$p<0.05$; ns $p>0.05$.

Figure 6 - HuR, hnRNP C and PTBP1 bind to the USE sequence in a specific manner.

(A) Schematic representation of the pre-cleaved pre-mRNAs used to perform the UV cross-linking assays. The initial black boxes correspond to a 39-nucleotides polylinker sequence present in the pre-mRNA. The light green boxes represent the USE sequence and the USE sequence mutated in 26 nucleotides (USEmt). Point mutations in the USE necessary to generate the USEmt sequence are represented in red, in the lower panel. AUUAAA corresponds to the proximal poly(A) signal. The cleaved pre-mRNAs were obtained by linearizing the plasmids with restriction enzymes in proximity to the cleavage site (portrayed by the pair of scissors). **(B)** UV cross-linking assay using the pre-cleaved pre-mRNAs USE and USEmt incubated with HeLa cells nuclear extracts. Asterisks define the bands correspondent to the proteins that undergo a decrease in the binding upon the mutation of the USE sequence to USEmt. **(C)** UV cross-linking competition assay using pre-cleaved pre-mRNAs USE and USEmt with HeLa cell nuclear extracts (lane 1) and an increasing amount of unlabeled USE (lane 2-4) and USEmt (lane 5-7) pre-mRNAs. One (lane 2, 5), 50 (lane 3,6) and 150 pmols (lane 4,7) of unlabeled oligonucleotides were used to compete with the labeled pre-mRNAs. **(D-E)** Immunoprecipitations assays performed subsequently to UV cross-linking of pre-cleaved pre-mRNAs USE and USEmt with HeLa cell nuclear extracts. Immunoprecipitations were performed using HuR (left panel) and hnRNP C (right panel) monoclonal antibodies.

Figure 7- PTBP1 knockdown affects the mRNA levels of human USE-containing genes. PTBP1 knockdown was performed using siPTBP1 and USE-containing genes mRNA levels were assessed by RT-qPCR. A siNTC pool was

also transfected to be used as a control. *EXT2* (**B**), *CLDN12* (**C**), *MED18* (**E**) and *DKC1* (**H**) mRNA levels decreased upon *PTBP1* knockdown, like expected, but *MBNL2* (D) and *CSNK1A1* (**G**) mRNA levels increased. For *KIF20A* (**A**), *NR5A2* (**F**) and *MAP4K5* (**I**) the difference was not statistically significant. The data is presented as the mean ± SD of at least three independent experiments analyzed by two-tailed unpaired Student's t-test, *p < 0.05, **p < 0.01, ***p < 0.001 and ns – not significant.

Figure 8 - A SNP in *C11orf53* generates an ectopic consensus of the USE which is associated with malignant tumor of colon

(A) Workflow used to identify SNPs that overlap with USE sequence and that had an associated disease. **(B)** USE related SNPs with the respective gene, disease, sequence impact and normal sequence. (**C**) Representative confocal images of microinjected embryos with GFP-USE of C11orf53 with and without the presence of the risk variant, showing GFP expression in gut cells of 48hpf zebrafish embryos. The dashed lines represent different GFP expression domain within an embryo(D) Representative graph of the quantification of the GFP expression in the defined GFP expression domains (n=14, normal variant; n=20 risk variant). The images were acquired using the confocal microscope LeicaSP5II and the GFP expression quantified using IMARIS software. The statistical analysis was done by two-tailed unpaired Student's t-test, *p < 0.05, **p < 0.01, ***p < 0.001.

Supplementary Table 1. Oligonucleotides and antibodies used in this study.

Figure S1. USE-containing genes GO terms from human (A), mouse (B) and zebrafish (C). The correspondent number of genes, fold enrichment and FDR are described for each GO term.

Figure S2. Microinjection of *USE* RNA leads to the decrease in GFP expression, development of a patchy notochord and eye defects in transgenic GFP-USE animals the appearance of eye defects, at 48 hours

post-fertilization. (A) Representative image of GFP expression in transgenic GFP-USE zebrafish embryos microinjected with *USE* RNA or *USEmt* RNA, at 48 hpf. Images were acquired using Leica M205Scale bar = 100 µm. **(B)** GFP expression was quantified in the notochord (n=37 USE;n=57 USEmt.. Determination of the percentage of patchy GFP expression in notochord (68%,n=79 USE; 22%,n=61 USEmt) **(C)**, eye size (n=53 USE; n=47 USEmt **(D)** and microphthalmy (28%,n=83 USE; 9%,n=91 USEmt) **(E)** of transgenic GFP-USE embryos upon the microinjection of *USE* RNA and *USEmt* RNA, at 48 hpf. Statistical significance was determined by two-tailed unpaired t-test and χ^2 test with Fisher correction. ***p<0.001; **p<0.01; ns p>0.05.

Figure S3. Microinjection of *GFP-USE* RNA in one-cell stage embryos leads to the development of various abnormalities 5 days upon the microinjection procedure.

(A) Quantification of the several defects developed 5 days upon the microinjection of *GFP-USE* RNA (in green) and *GFP-USEmt* RNA (in grey) in one-cell stage zebrafish embryos. Statistical significance was determined using χ^2 with Fisher correction. **(B)** Pie chart representing the total number of defects developed upon the microinjection of *GFP-USE* RNA and *GFP-USEmt* RNA. **(C)** Representative images showing the various abnormalities upon microinjection of *GFP-USE* RNA at 5 dpf, indicated with red arrows. Scale bar = 100 µm. Statistical significance was determined by χ^2 test with Fisher correction or by two-tailed unpaired t-test. ***p<0.001; **p<0.01; *p<0.05; ns p>0.05 Images acquired with Leica M205.Scale bar = 100 µm

Figure S4. Mapping the binding sites for hnRNP C and HuR.

Representation of the USE sequence and the point mutations that were introduced to generate mutant sequences (USEmt-2; USE-5'mt; USE-3'mut). Representation of short RNAs USE-(a); USE-(b); USE-(c) generated in order to map the RBPs'

binding sites. **(B)** UV cross-linking using pre-cleaved pre-mRNAs USEmt-2, USE-5'mt and USE-3'mt, with HeLa cell nuclear extracts. **(C)** UV cross-linking using the short RNAs- USE-(a), USE-(b) and USE-(c)- and HeLa cell nuclear extracts. pGEM7 (lane 5) corresponds to an RNA transcribed from a linearized polylinker, acting as a negative control.

Figure S5. *PTBP1* gene expression was efficiently silenced in HeLa cells.

In order to knockdown *PTBP1*, HeLa cells were transfected with two siRNAs targeting *PTBP1*. A siNTC was also transfected, to act as a negative control. mRNA levels of *PTBP1* were assessed by RT-qPCR, showing a mean knockdown of ~75%.

References

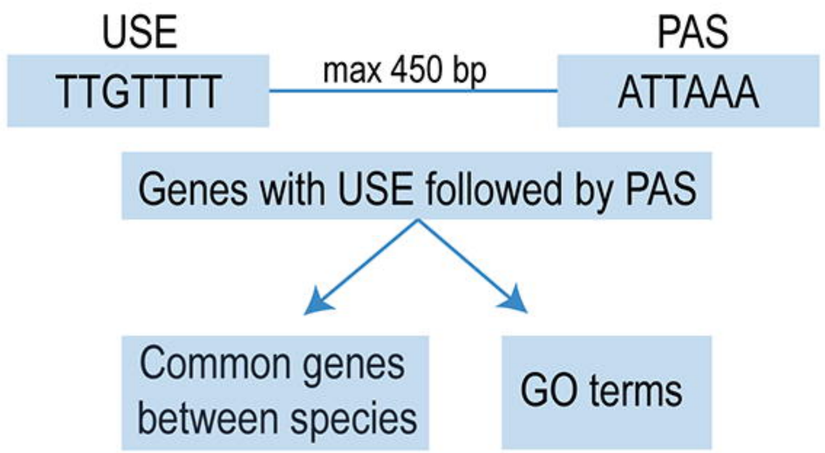
1. Mayr, C., *Regulation by 3'-Untranslated Regions.* Annu Rev Genet, 2017. **51**: p. 171-194.
2. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
3. Tian, B. and J.L. Manley, *Alternative polyadenylation of mRNA precursors.* Nat Rev Mol Cell Biol, 2017. **18**(1): p. 18-30.
4. Szostak, E. and F. Gebauer, *Translational control by 3'-UTR-binding proteins.* Brief Funct Genomics, 2013. **12**(1): p. 58-65.
5. Mayr, C., *What Are 3' UTRs Doing?* Cold Spring Harb Perspect Biol, 2019. **11**(10).
6. Pereira-Castro, I. and A. Moreira, *On the function and relevance of alternative 3'-UTRs in gene expression regulation.* Wiley Interdiscip Rev RNA, 2021. **12**(5): p. e1653.
7. Tian, B. and J.H. Graber, *Signals for pre-mRNA cleavage and polyadenylation.* Wiley Interdiscip Rev RNA, 2012. **3**(3): p. 385-96.
8. Ulitsky, I., et al., *Extensive alternative polyadenylation during zebrafish development.* Genome Res, 2012. **22**(10): p. 2054-66.
9. Beaudoing, E., et al., *Patterns of variant polyadenylation signal usage in human genes.* Genome Res, 2000. **10**(7): p. 1001-10.
10. Tian, B., et al., *A large-scale analysis of mRNA polyadenylation of human and mouse genes.* Nucleic Acids Res, 2005. **33**(1): p. 201-12.
11. Moreira, A., et al., *The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms.* Genes Dev, 1998. **12**(16): p. 2522-34.
12. Oliveira, M.S., et al., *Cell Cycle Kinase Polo Is Controlled by a Widespread 3' Untranslated Region Regulatory Sequence in Drosophila melanogaster.* Mol Cell Biol, 2019. **39**(15).
13. Darmon, S.K. and C.S. Lutz, *Novel upstream and downstream sequence elements contribute to polyadenylation efficiency.* RNA Biol, 2012. **9**(10): p. 1255-65.
14. Kaufmann, I., et al., *Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase.* Embo j, 2004. **23**(3): p. 616-26.
15. Danckwardt, S., et al., *p38 MAPK controls prothrombin expression by regulated RNA 3' end processing.* Mol Cell, 2011. **41**(3): p. 298-310.
16. Curinha, A., et al., *Implications of polyadenylation in health and disease.* Nucleus, 2014. **5**(6): p. 508-19.
17. Nedeljkovic, M., et al., *Long-distance regulation of Add2 pre-mRNA3'end processing.* RNA Biol, 2013. **10**(4): p. 516-27.
18. Lutz, C.S. and A. Moreira, *Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression.* Wiley Interdiscip Rev RNA, 2011. **2**(1): p. 23-31.

19. Natalizio, B.J., et al., *Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals.* J Biol Chem, 2002. **277**(45): p. 42733-40.
20. Pinto, P.A., et al., *RNA polymerase II kinetics in polo polyadenylation signal selection.* Embo j, 2011. **30**(12): p. 2431-44.
21. Team, R.C. *R: A language and environment for statistical computing*. 2018; Available from: https://www.R-project.org/.
22. Kawakami, K., *Tol2: a versatile gene transfer vector in vertebrates.* Genome Biol, 2007. **8 Suppl 1**(Suppl 1): p. S7.
23. Richardson, R., et al., *The zebrafish eye-a paradigm for investigating human ocular genetics.* Eye (Lond), 2017. **31**(1): p. 68-86.
24. López de Silanes, I., et al., *Identification of a target RNA motif for RNA-binding protein HuR.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 2987-92.
25. Myer, V.E., X.C. Fan, and J.A. Steitz, *Identification of HuR as a protein implicated in AUUUA-mediated mRNA decay.* Embo j, 1997. **16**(8): p. 2130-9.
26. Ma, W.J., et al., *Cloning and characterization of HuR, a ubiquitously expressed Elav-like protein.* J Biol Chem, 1996. **271**(14): p. 8144-51.
27. Ray, D., et al., *A compendium of RNA-binding motifs for decoding gene regulation.* Nature, 2013. **499**(7457): p. 172-7.
28. Hinman, M.N. and H. Lou, *Diverse molecular functions of Hu proteins.* Cell Mol Life Sci, 2008. **65**(20): p. 3168-81.
29. Barnett, S.F., D.L. Friedman, and W.M. LeStourgeon, *The C proteins of HeLa 40S nuclear ribonucleoprotein particles exist as anisotropic tetramers of (C1)3 C2.* Mol Cell Biol, 1989. **9**(2): p. 492-8.
30. Geuens, T., D. Bouhy, and V. Timmerman, *The hnRNP family: insights into their role in health and disease.* Hum Genet, 2016. **135**(8): p. 851-67.
31. Görlach, M., C.G. Burd, and G. Dreyfuss, *The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins.* J Biol Chem, 1994. **269**(37): p. 23074-8.
32. Hamilton, B.J., et al., *Association of heterogeneous nuclear ribonucleoprotein A1 and C proteins with reiterated AUUUA sequences.* J Biol Chem, 1993. **268**(12): p. 8881-7.
33. Linares, A.J., et al., *The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation.* Elife, 2015. **4**: p. e09268.
34. Lou, H., R.F. Gagel, and S.M. Berget, *An intron enhancer recognized by splicing factors activates polyadenylation.* Genes Dev, 1996. **10**(2): p. 208-19.
35. Sawicka, K., et al., *Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein.* Biochem Soc Trans, 2008. **36**(Pt 4): p. 641-7.
36. Castelo-Branco, P., et al., *Polypyrimidine tract binding protein modulates efficiency of polyadenylation.* Mol Cell Biol, 2004. **24**(10): p. 4174-83.
37. Wilusz, J. and T. Shenk, *A uridylate tract mediates efficient heterogeneous nuclear ribonucleoprotein C protein-RNA cross-linking and functionally substitutes for the downstream element of the polyadenylation signal.* Mol Cell Biol, 1990. **10**(12): p. 6397-407.

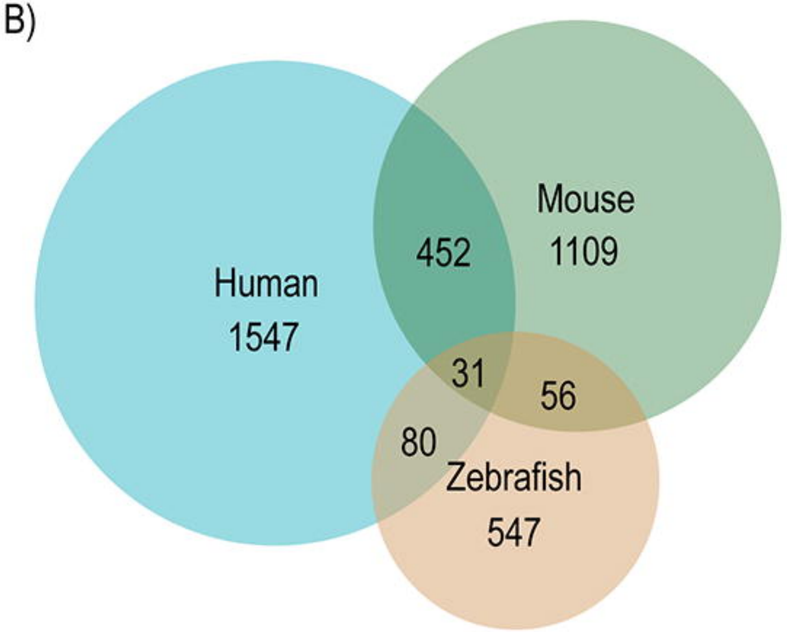
38. Lee, T.I. and R.A. Young, *Transcriptional regulation and its misregulation in disease.* Cell, 2013. **152**(6): p. 1237-51.
39. Harris, B.T., et al., *Transcriptional dynamics of colorectal cancer risk associated variation at 11q23.1 correlate with tuft cell abundance and marker expression in silico.* Sci Rep, 2022. **12**(1): p. 13609.
40. Liu, Y., et al., *Genome-Wide Study Links PNPLA3 Variant With Elevated Hepatic Transaminase After Acute Lymphoblastic Leukemia Therapy.* Clin Pharmacol Ther, 2017. **102**(1): p. 131-140.
41. Szczepanski, A.P., et al., *POU2AF2/C11orf53 functions as a coactivator of POU2F3 by maintaining chromatin accessibility and enhancer activity.* Sci Adv, 2022. **8**(40): p. eabq2403.
42. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.
43. Kinsella, R.J., et al., *Ensembl BioMarts: a hub for data retrieval across taxonomic space.* Database (Oxford), 2011. **2011**: p. bar030.
44. Mi, H., et al., *PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools.* Nucleic Acids Res, 2019. **47**(D1): p. D419-d426.
45. Piñero, J., et al., *The DisGeNET knowledge platform for disease genomics: 2019 update.* Nucleic Acids Res, 2020. **48**(D1): p. D845-d855.
46. Llamazares, S., et al., *polo encodes a protein kinase homolog required for mitosis in Drosophila.* Genes Dev, 1991. **5**(12a): p. 2153-65.
47. Karlsson, J., J. von Hofsten, and P.E. Olsson, *Generating transparent zebrafish: a refined method to improve detection of gene expression during embryonic development.* Mar Biotechnol (NY), 2001. **3**(6): p. 522-7.
48. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis.* Nat Methods, 2012. **9**(7): p. 676-82.
49. McCloy, R.A., et al., *Partial inhibition of Cdk1 in G 2 phase overrides the SAC and decouples mitotic events.* Cell Cycle, 2014. **13**(9): p. 1400-12.
50. Pfaffl, M.W., *A new mathematical model for relative quantification in real-time RT-PCR.* Nucleic Acids Res, 2001. **29**(9): p. e45.
51. El-Bazzal, L., et al., *A novel EXT2 mutation in a consanguineous family with severe developmental delay, microcephaly, seizures, feeding difficulties, and osteopenia extends the phenotypic spectrum of autosomal recessive EXT2-related syndrome (AREXT2).* Eur J Med Genet, 2019. **62**(4): p. 259-264.
52. Alghamdi, R.A. and M.H. Al-Zahrani, *Identification of key claudin genes associated with survival prognosis and diagnosis in colon cancer through integrated bioinformatic analysis.* Front Genet, 2023. **14**: p. 1221815.
53. Rajagopalan, L.E., et al., *hnRNP C increases amyloid precursor protein (APP) production by stabilizing APP mRNA.* Nucleic Acids Res, 1998. **26**(14): p. 3418-23.
54. Brennan, C.M. and J.A. Steitz, *HuR and mRNA stability.* Cell Mol Life Sci, 2001. **58**(2): p. 266-77.
55. Rothamel, K., et al., *ELAVL1 primarily couples mRNA stability with the 3' UTRs of interferon-stimulated genes.* Cell Rep, 2021. **35**(8): p. 109178.

879

A)



B)

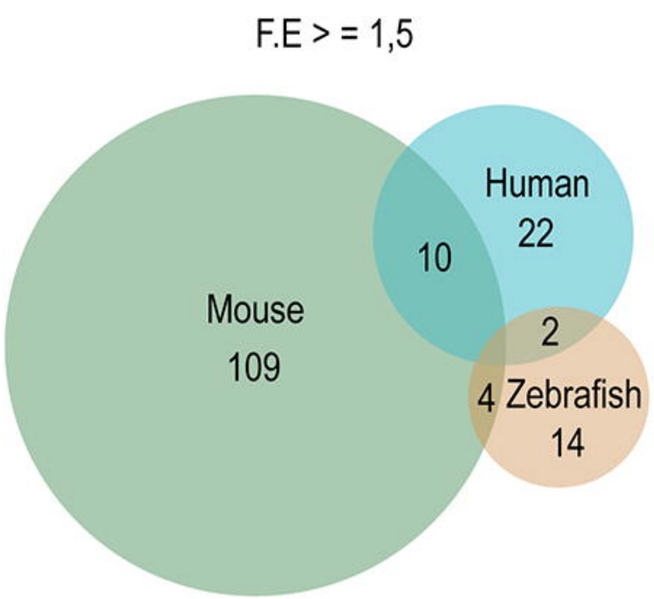


C)

31 USE-containing genes common in Human, Mouse and Zebrafish

ACBD3	*MED18*	*DCLRE1B*	*RNF180*
CAND1	*NEGR1*	*DKC1*	*SOD2*
CDKN2C	*NPEPPS*	*DYNLT3*	*STRN*
CHUK	*NR5A2*	*EXT2*	*STX6*
CLDN12	*PPM1K*	*KIF20A*	*UBE3C*
CNPY1	*PRDM10*	*LUC7L*	*YES1*
CPSF6	*RAB27A*	*MAP4K5*	*ZMAT2*
CSNK1A1	*RGMB*	*MBNL2*	

D)

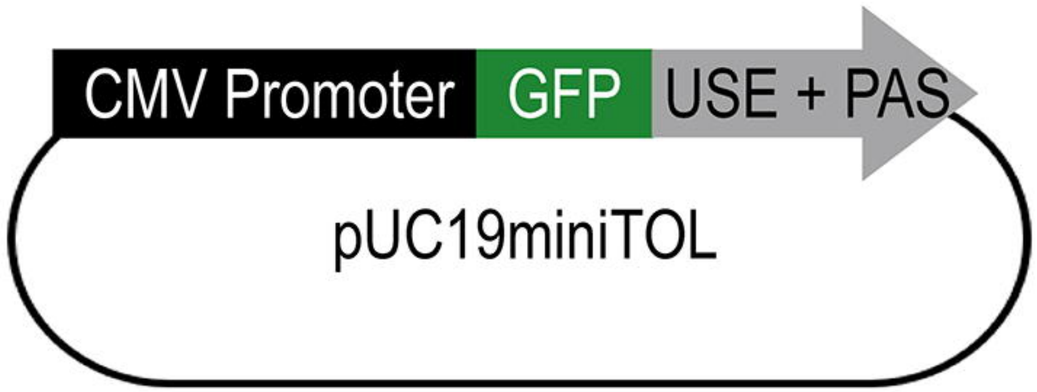


E)

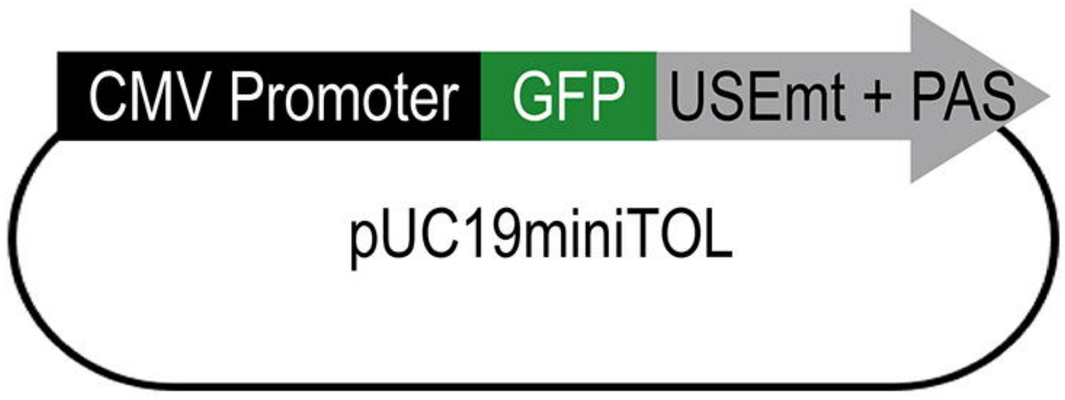
	Human	Mouse	Zebrafish
positive regulation of cell migration (GO:0030335)	•	•	
mRNA processing (GO:0006397)	•	•	
positive regulation of cell motility (GO:2000147)	•	•	
response to growth factor (GO:0070848)	•	•	
positive regulation of locomotion (GO:0040017)	•	•	
cell division (GO:0051301)	•	•	
positive regulation of cellular component movement(GO:0051272)	•	•	
cellular response to growth factor stimulus (GO:0071363)	•	•	
chromatin organization (GO:0006325)	•	•	
mRNA metabolic process (GO:0016071)	•	•	
chromosome organization (GO:0051276)		•	•
establishment of localization in cell (GO:0051649)		•	•
cellular protein localization (GO:0034613)		•	•
cellular macromolecule localization (GO:0070727)		•	•
mitotic cell cycle (GO:0000278)	•		•
mitotic cell cycle process (GO:1903047)	•		•

A)

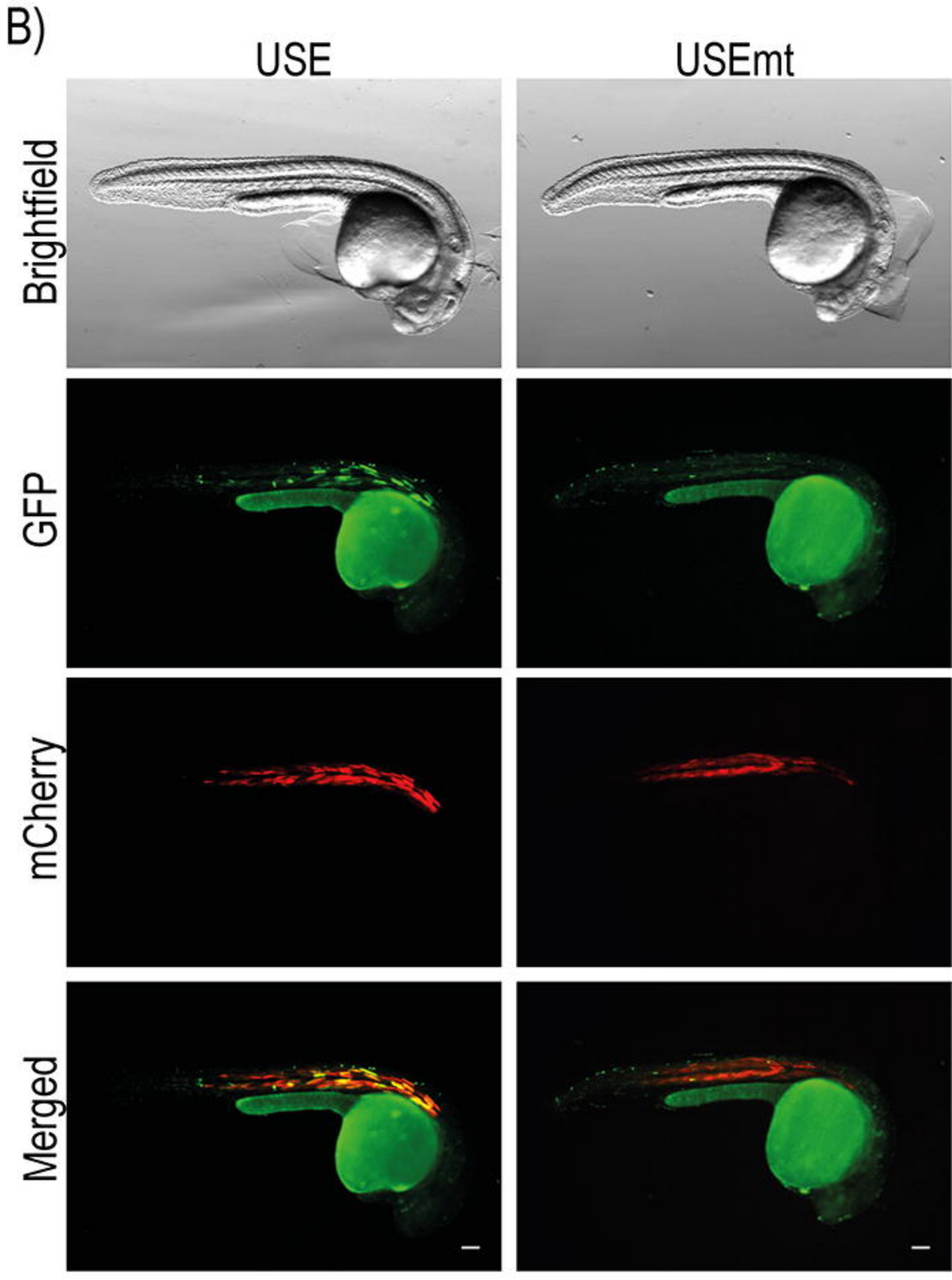
USE 5`- ATTTGTTTTTGCCCCTTCCCCTT - 3`



USEmt 5`- ATAGACGACAAACGCAAGATCTT - 3`

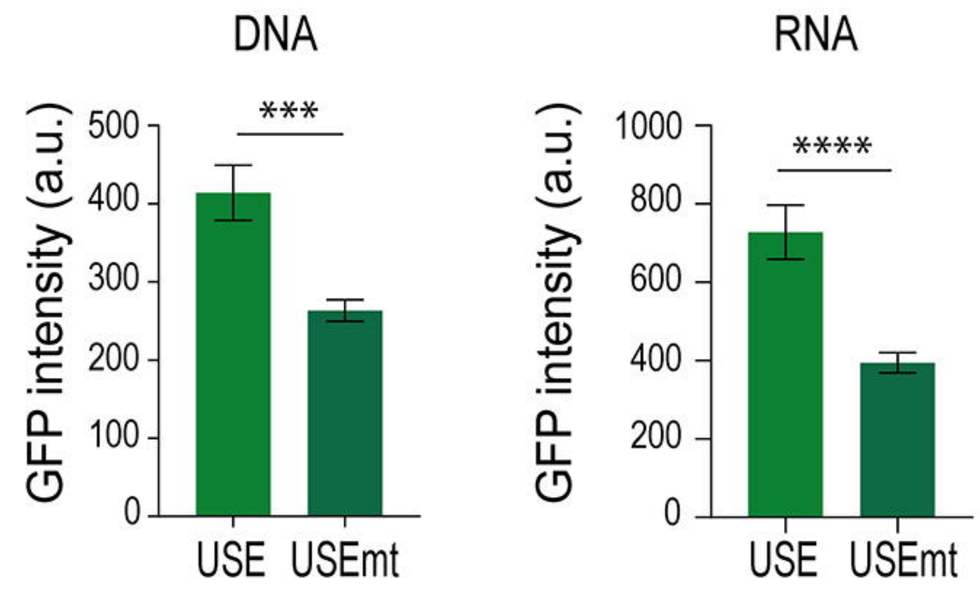


B)

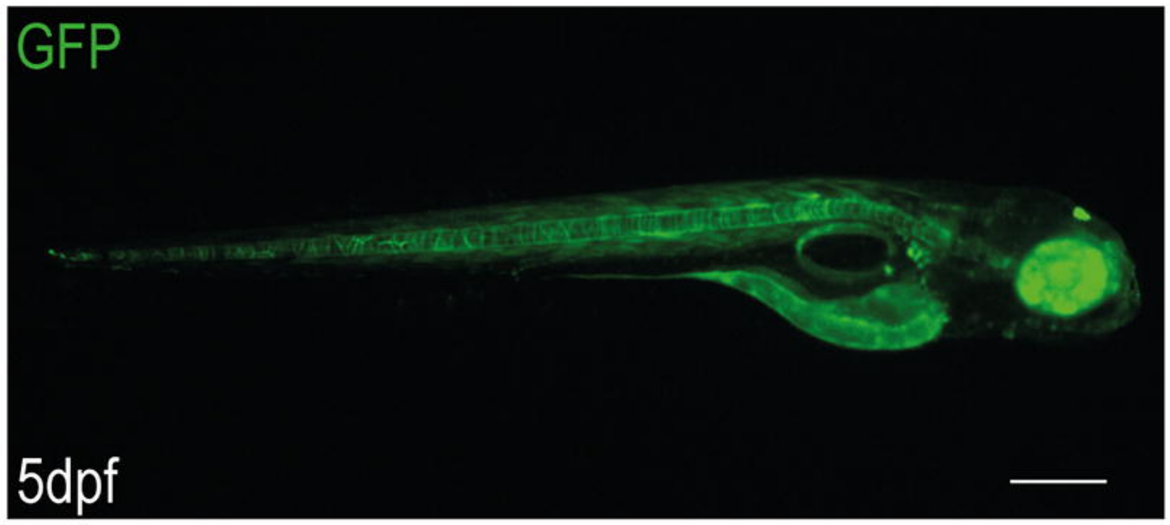


C)

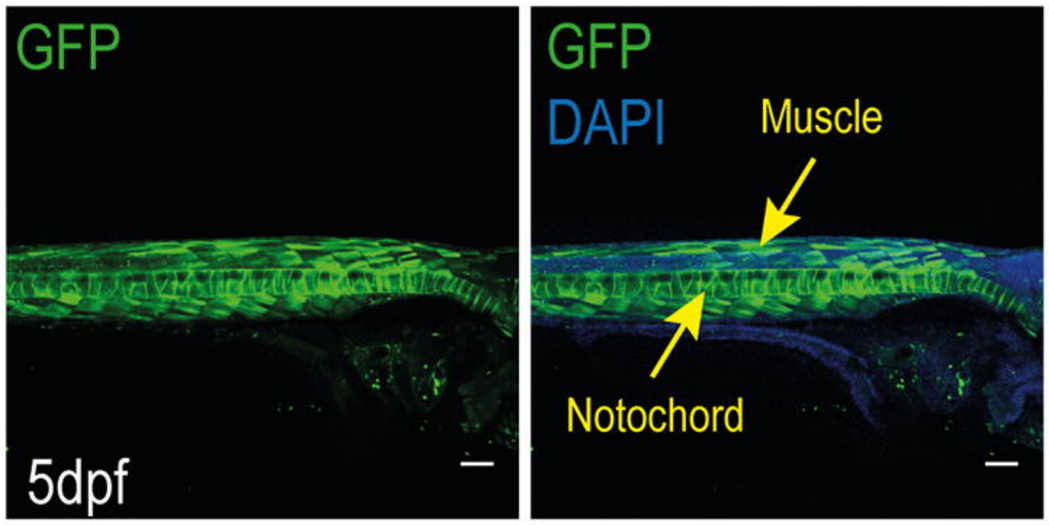
D)



A)



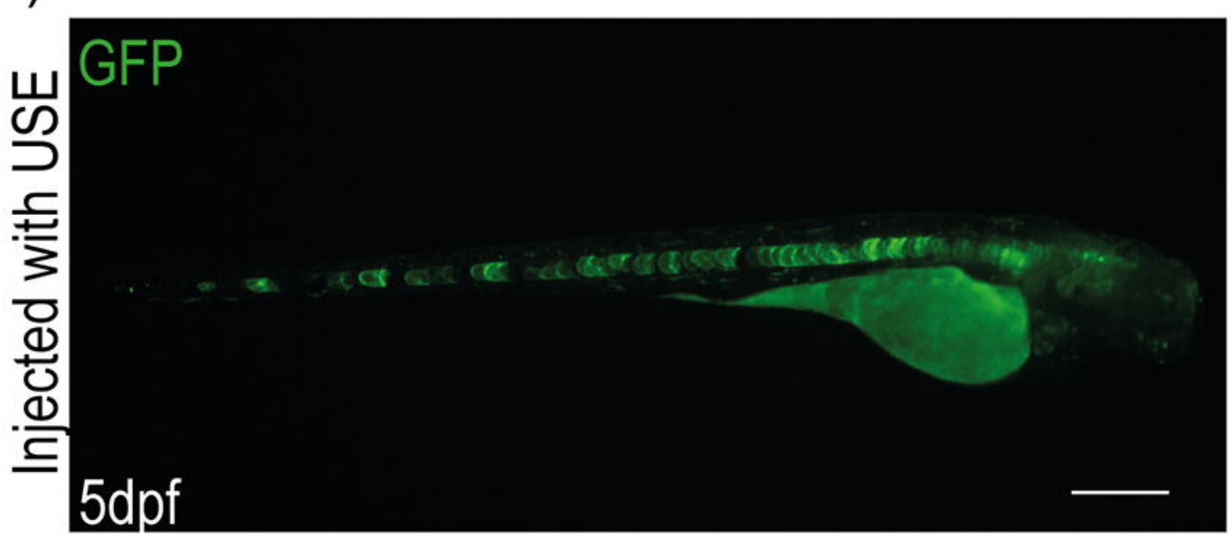
B)

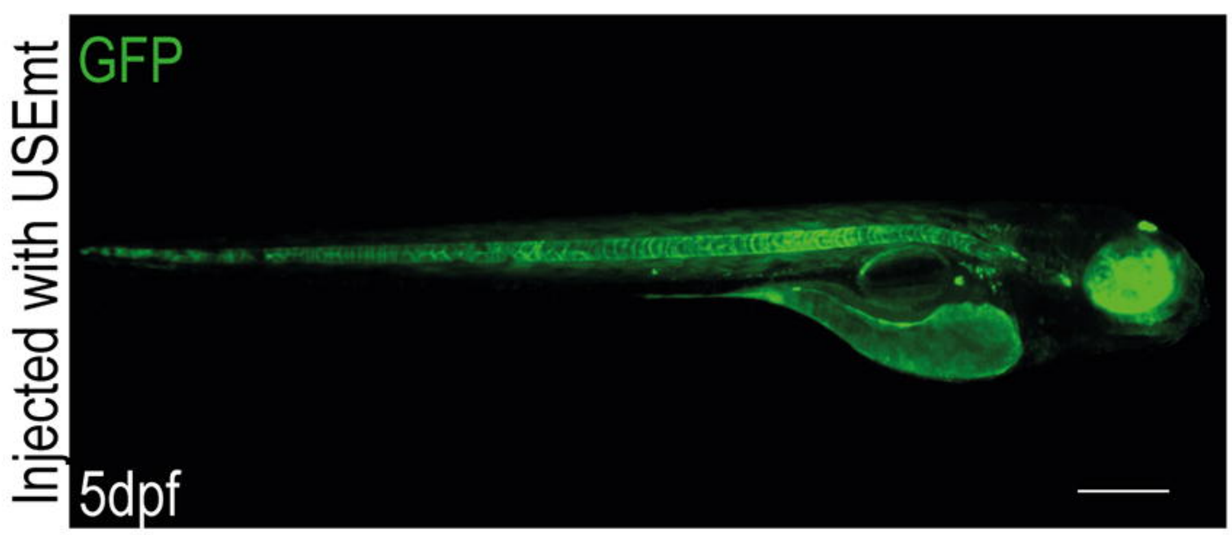


C)

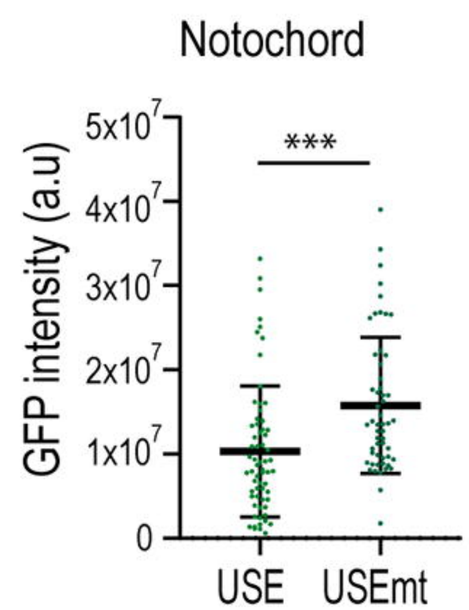
USE	5'-AAUUUAUUUGUUUUUGCCCCUUCCCCCUU-3'
USEmt	5'-AAUUCAUAGACGACAAACGCAAGAUCUU-3'

D)

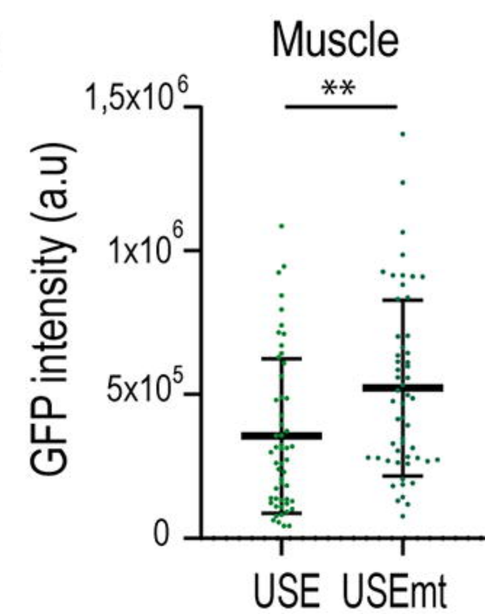




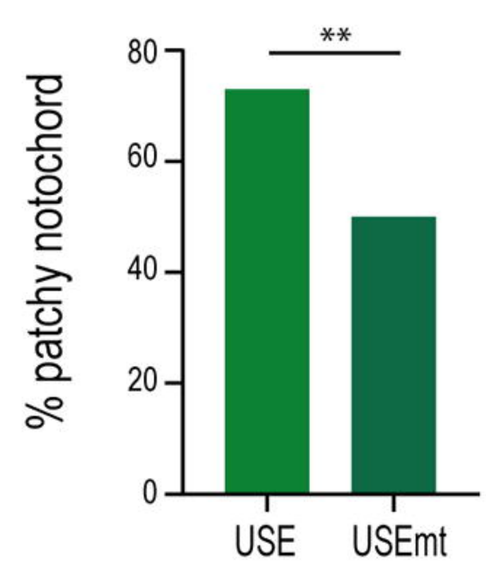
E)



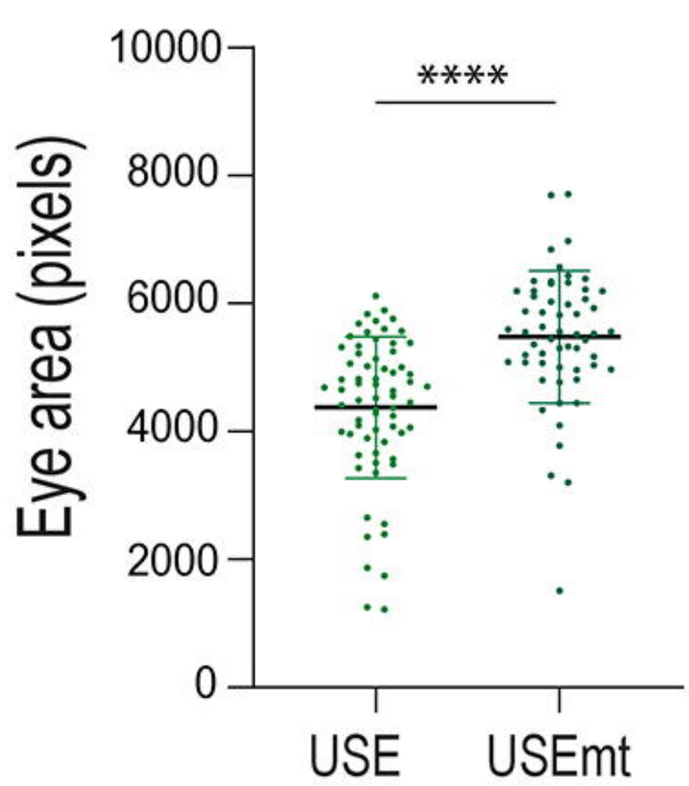
F)



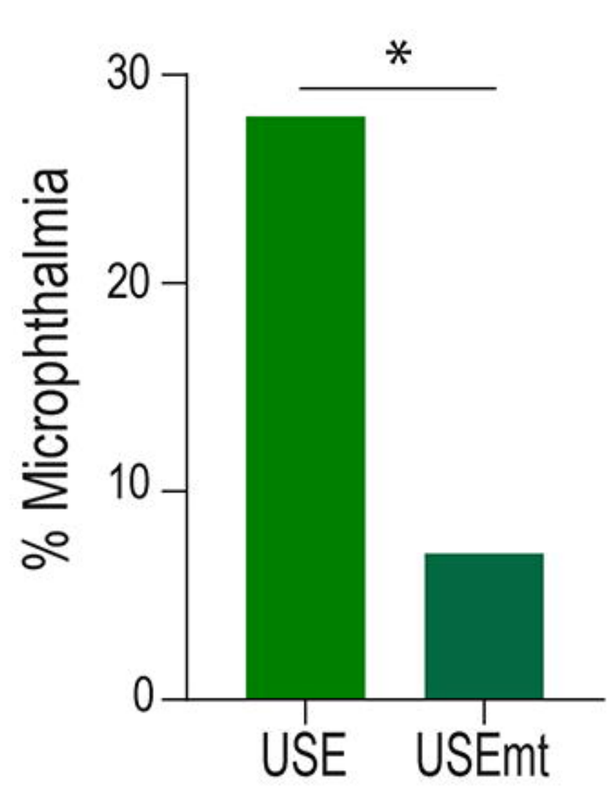
G)



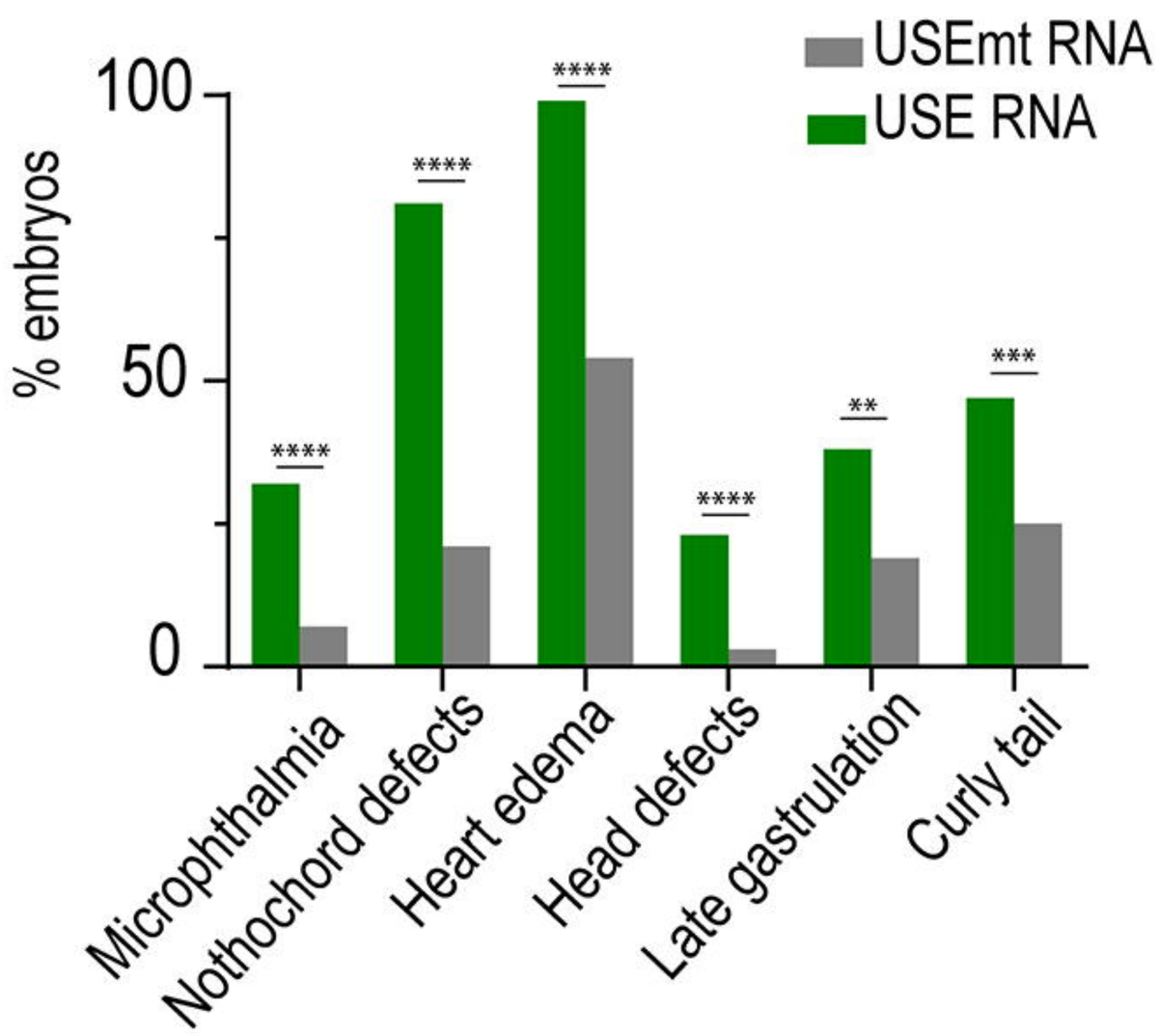
A)



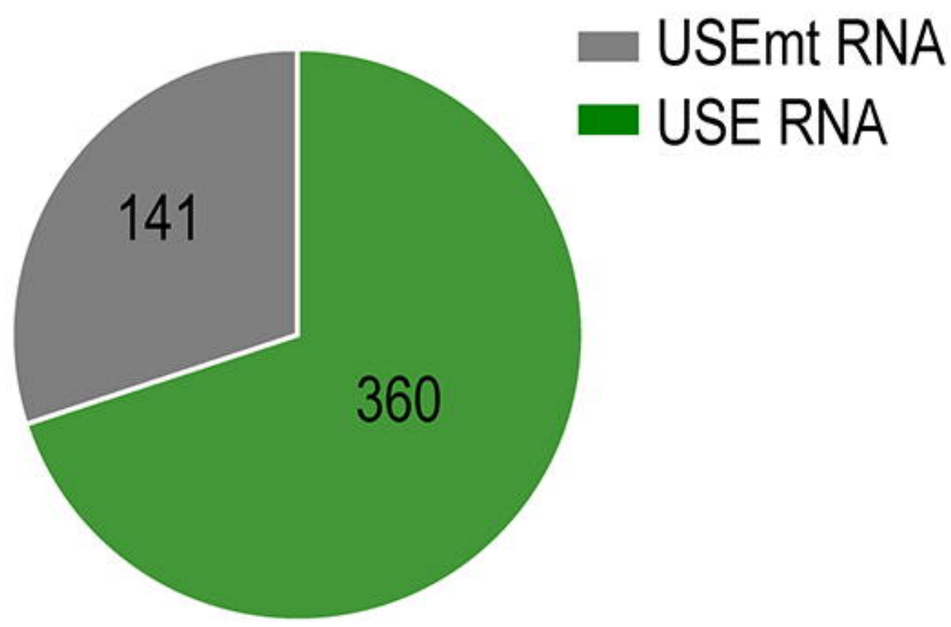
B)



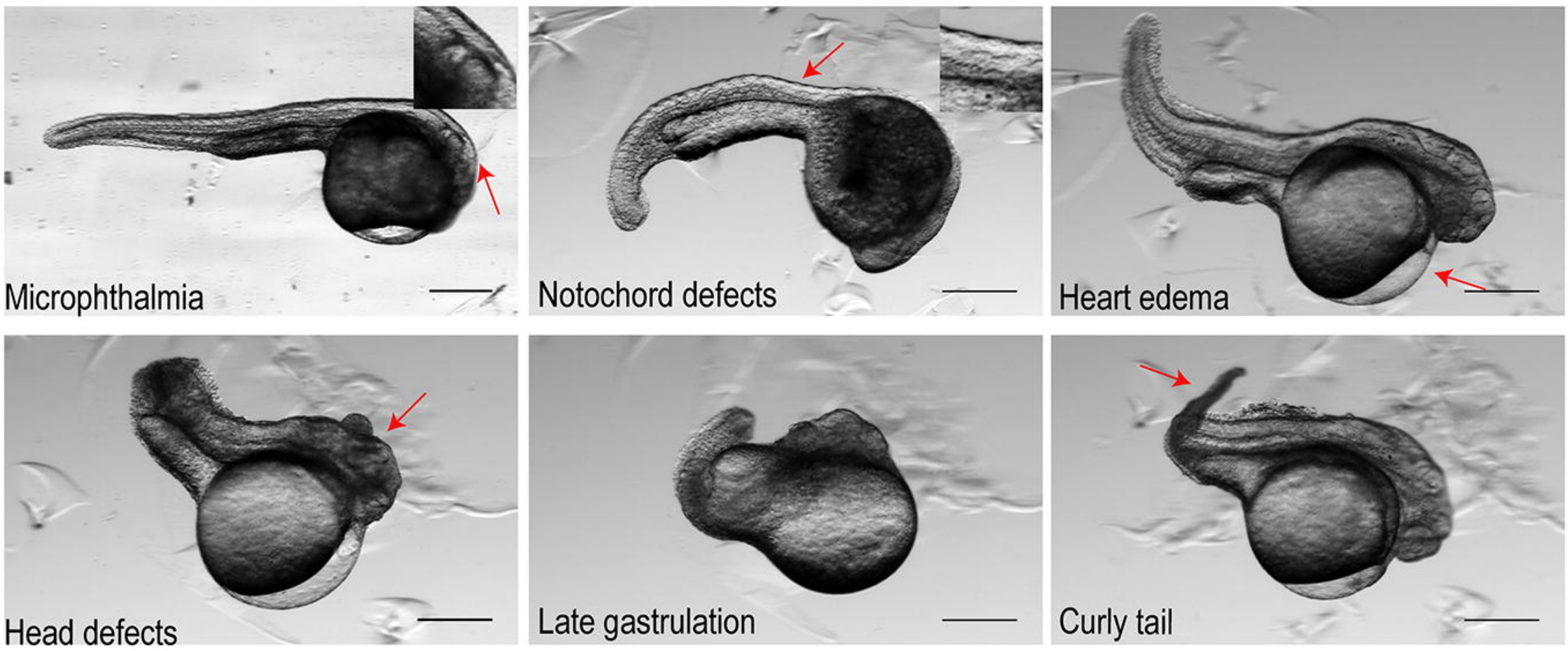
C) Defects on 24hpf zebrafish embryos



D) Total number of defects in 24hpf zebrafish embryos

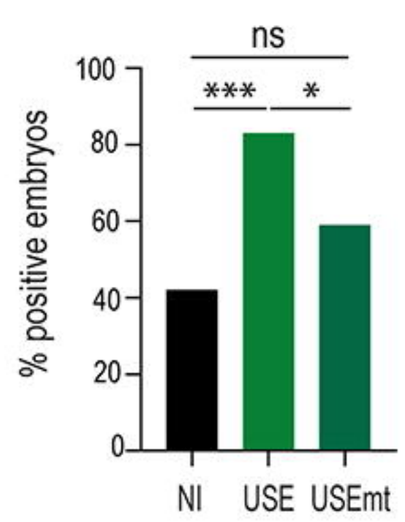


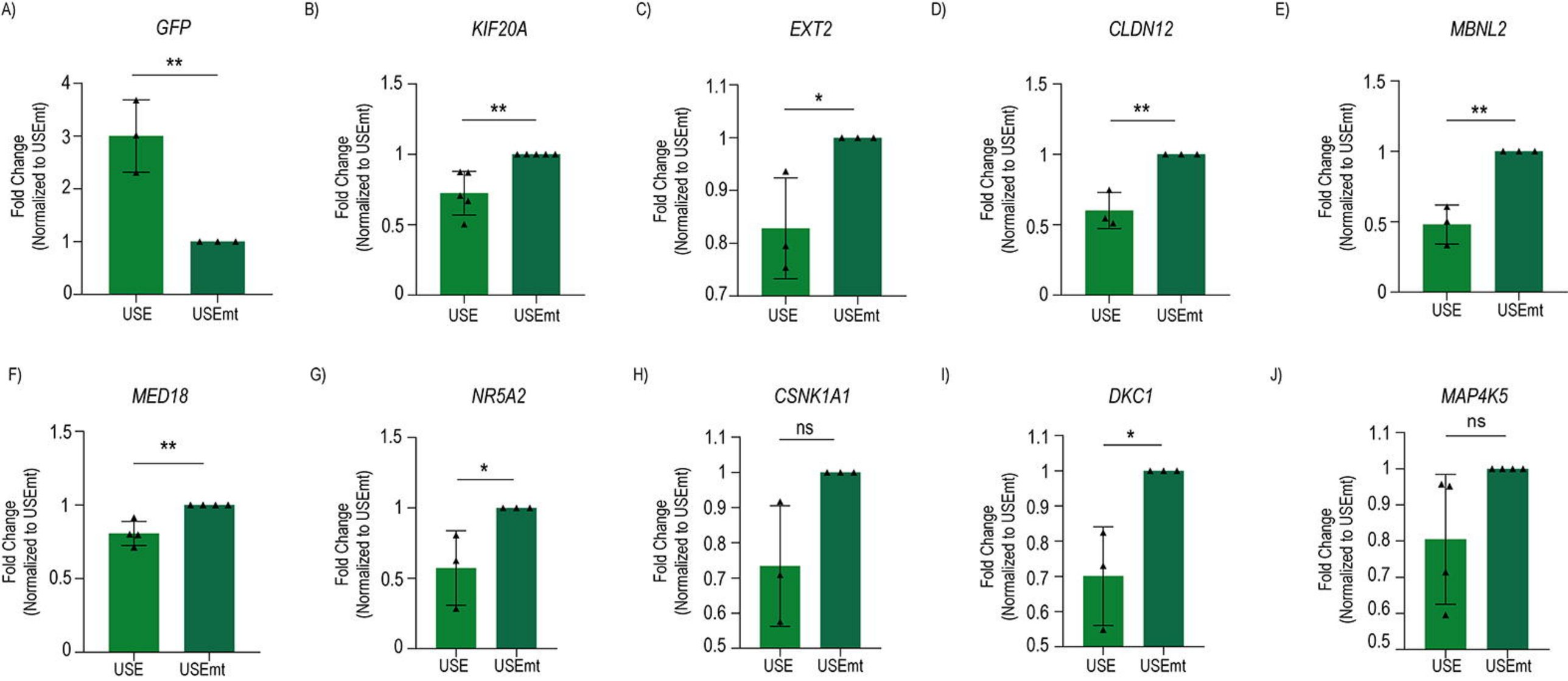
E)



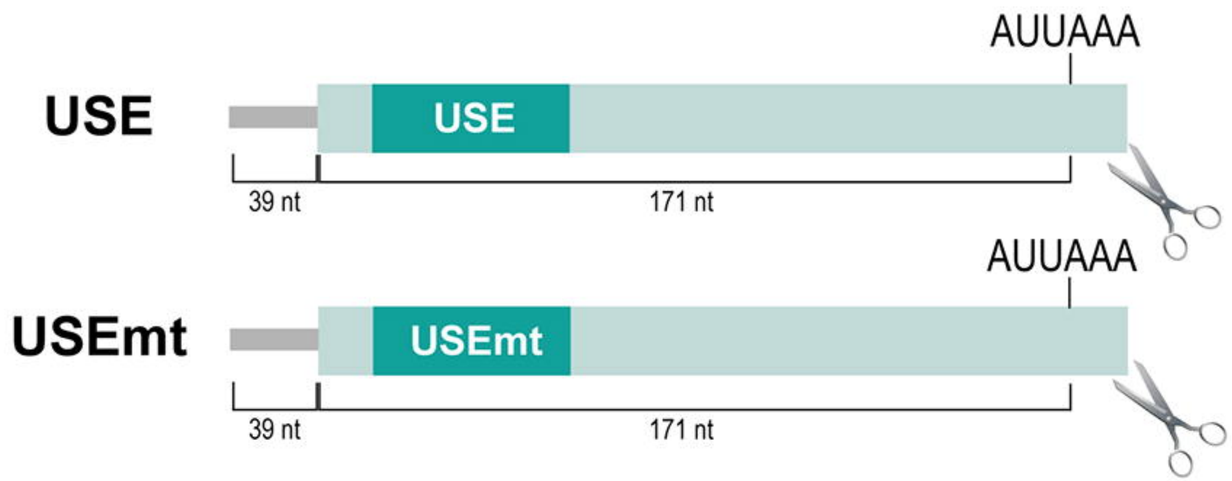
F)

G)



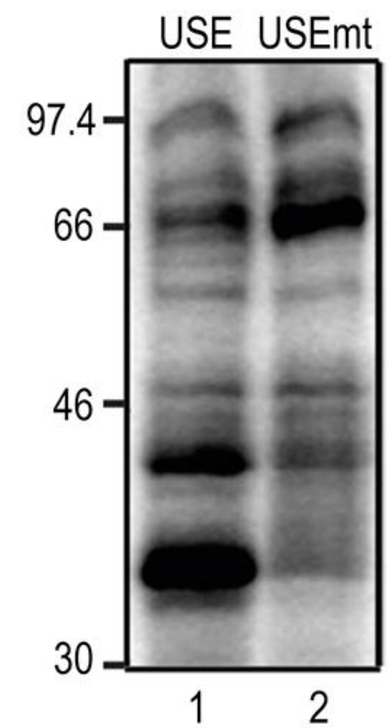

A)
GFP
**
Fold Change (Normalized to USEmt)
4
3
2
1
0
USE
USEmt
B)
KIF20A
**
1.5
1
0.5
0
C)
EXT2
*
1.1
1
0.9
0.8
0.7
D)
CLDN12
**
E)
MBNL2
**
F)
MED18
**
G)
NR5A2
*
H)
CSNK1A1
ns
I)
DKC1
*
J)
MAP4K5
ns
0.6
0.5

A)

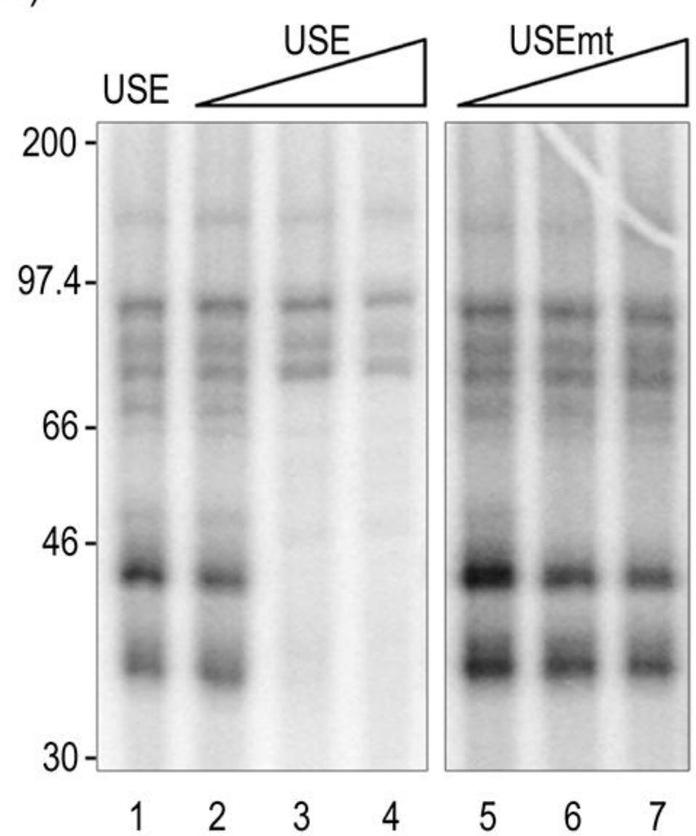


USE	5'-AAUUUAUUUGUUUUUGCCCCUUCCCCUU-3'
USEmt	5'-UUAAAUAAACAGAAACGGAGAAGGAAUU-3'

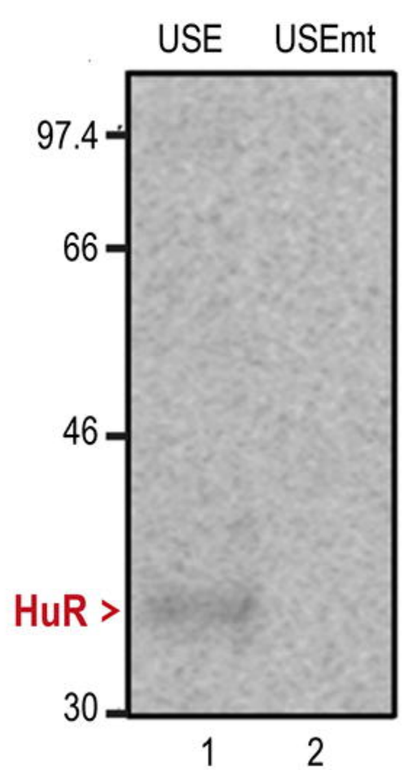
B)



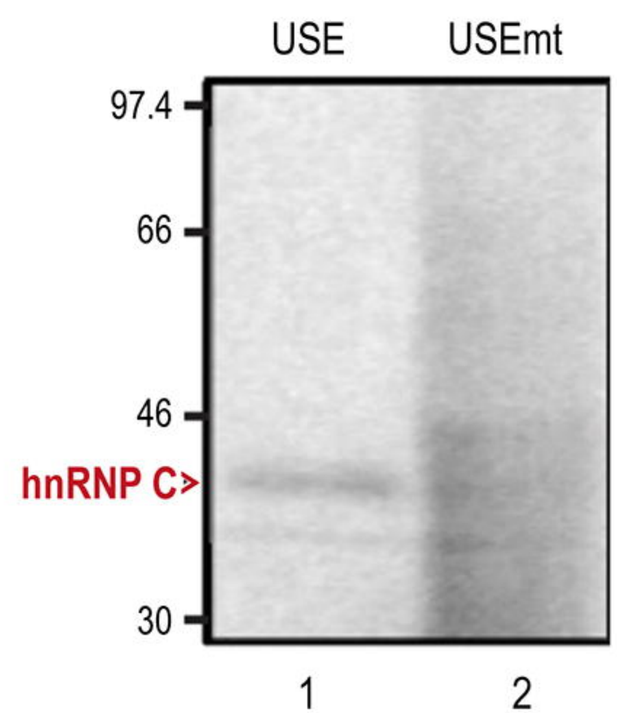
C)



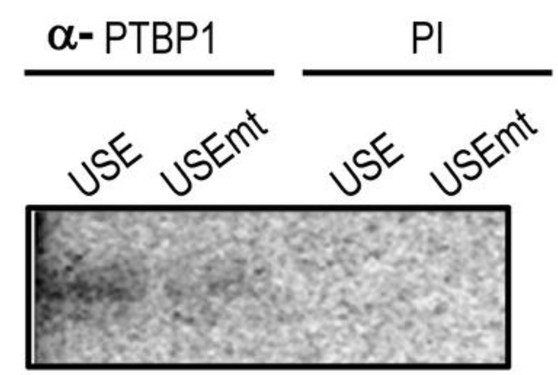
D)

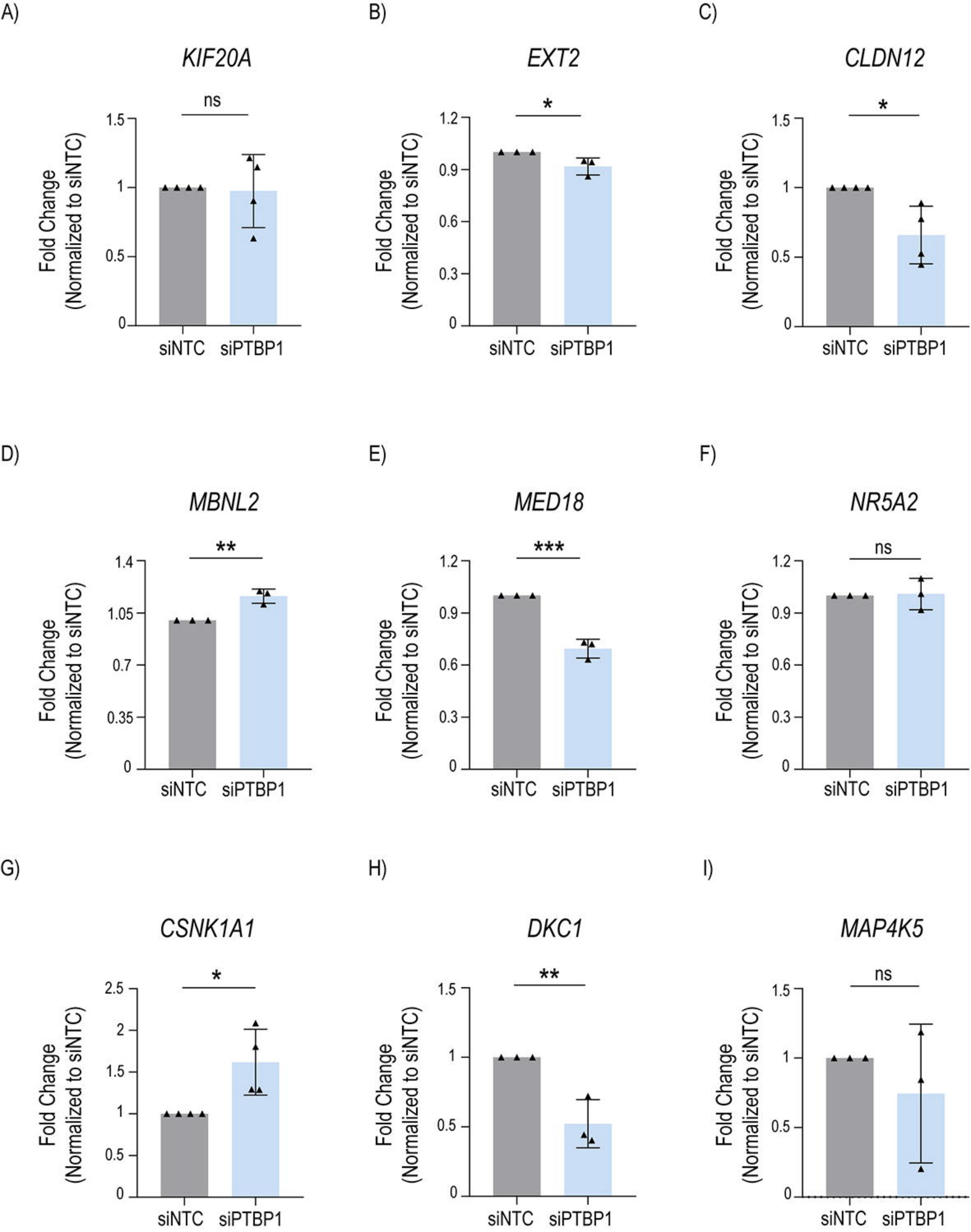


E)

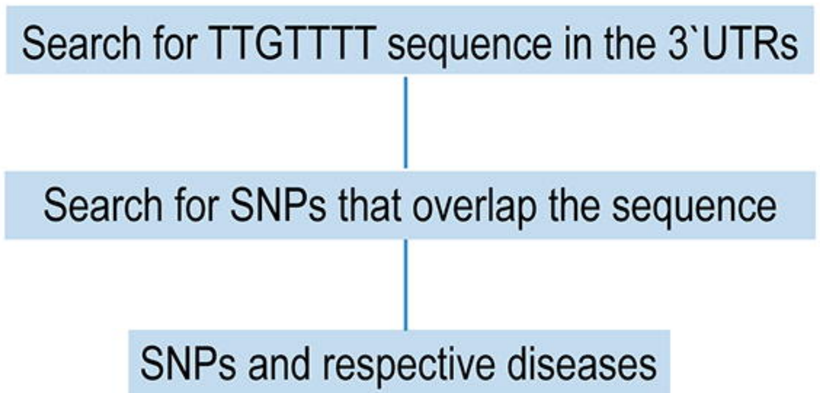


F)

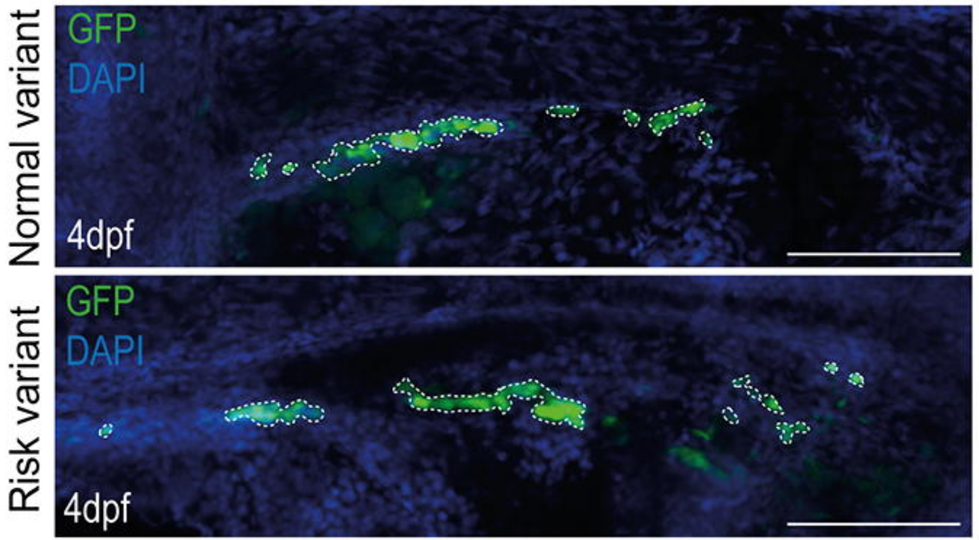



A)
KIF20A
ns
Fold Change (Normalized to siNTC)
1.5
1
0.5
0
siNTC
siPTBP1
B)
EXT2
*
1.2
0.9
0.6
0.3
0
C)
CLDN12
*
D)
MBNL2
**
1.4
1.05
0.7
0.35
E)
MED18
***
F)
NR5A2
ns
G)
CSNK1A1
*
2.5
2
1.5
1
0.5
0
H)
DKC1
**
I)
MAP4K5
ns

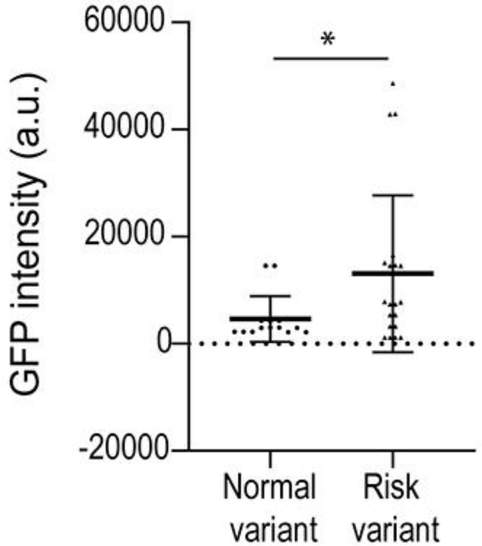
A)



C)



D)



B)

Gene	Disease	SNP	Risk variant	Sequence impact	Normal sequence
GREB1	Alanine aminotransferase measurement	rs149940960	T	TTGTTTT	TTGCTTT
C11orf53	Malignant tumor of colon	rs3087967	T	TTGTTTT	TCGTTTT