Mol Cell Proteomics Papers in Press. Published on September 4, 2018 as Manuscript RA118.000593

The protein coded by a short open reading frame, not by the annotated coding sequence is the main gene product of the dual-coding gene *MIEF1*

Vivian Delcourt[1,2,3], Mylène Brunelle[1,3], Annie V. Roy[1,3], Jean-François Jacques[1,3], Michel Salzet[2], Isabelle Fournier[2], Xavier Roucou[1,3]*

(1) Département de Biochimie, Université de Sherbrooke, Québec, Canada

(2) Univ. Lille, INSERM U1192, Laboratoire Protéomique, Réponse Inflammatoire & Spectrométrie de Masse (PRISM) F-59000 Lille, France

(3) PROTEO, Québec Network for Research on Protein Function, Structure, and Engineering, Québec, Canada

* Correspondance : xavier.roucou@usherbrooke.ca

Abstract

Proteogenomics and ribosome profiling concurrently show that genes may code for both a large and one or more small proteins translated from annotated coding sequences (CDSs) and unannotated alternative open reading frames (named alternative ORFs or altORFs), respectively, but the stoichiometry between large and small proteins translated from a same gene is unknown. MIEF1, a gene recently identified as a dual-coding gene, harbours a CDS and a newly annotated and actively translated altORF located in the 5'UTR . Here, we use absolute quantification with stable isotope-labeled peptides and parallel reaction monitoring to determine levels of both proteins in two human cells lines and in human colon. We report that the main MIEF1 translational product is not the canonical 463 amino acid MiD51 protein but the small 70 amino acid alternative MiD51 protein (altMiD51). These results demonstrate the inadequacy of the single CDS concept and provide a strong argument for incorporating altORFs and small proteins in functional annotations.

1 Introduction

According to the traditional view of protein synthesis, each protein-coding gene harbours a single annotated ORF or coding sequence (CDS) encoding a canonical protein. However, genes contain more than one ORF, and the longest ORF is generally designated as the canonical CDS in genome annotations [1]. In eukaryotes, alternative splicing results in the production of several mRNAs and the translation of different isoforms, in addition to the canonical protein. Hence, the translational output of a protein-coding gene is currently concealed to a canonical protein and one or several isoforms.

This concept was recently disproved by two modern approaches to the accurate measurement of translation, ribosome profiling and proteogenomics. Ribosome profiling maps the regions of the transcriptome which are actively translated with nucleotide resolution [2]. Proteogenomics approaches use customized protein databases and mass spectrometry (MS)-based proteomics to detect translated proteins [3–10]. Both methods have revealed prevalent translation of ORFs outside of annotated CDSs and of out-of-frame ORFs (altORFs) [2, 11]. These findings call into question the concept of the single CDS in eukaryotic mRNAs [12]. In addition, they also highlight the need to redefine translated sequences [13], modernize functional genome annotations with shorter ORFs [14], and reassess the translation output of protein coding-genes by considering smaller proteins in addition to larger canonical proteins. In particular, the cellular stoichiometry of a canonical protein versus a small protein encoded in the same gene and their respective concentrations is unknown. Yet, proteins are the primary effectors of biological processes and deciphering the function of a gene in health and disease requires accurate characterization of its products.

The mitochondrial elongation factor 1 gene or *MIEF1* also termed *SMCR7L/MiD51*, localized at the Chr22q13.1 locus, codes for a mitochondrial receptor of Drp1, a GTPase which functions in mitochondrial fission [15– 17].

Ribosome profiling and proteogeomics studies recently demonstrated the translation of a stable 70 amino acid protein product encoded in a altORF localized in the 5'UTR (Figure 1 a) [3, 18–24]. Thus, *MIEF1* is a prototypical gene coding for both a large and a small protein. For simplicity, we termed this novel protein "alternative-MiD51" or altMiD51. Remarkably, both proteins are localized at the mitochondria. MiD51 is an outer mitochondrial membrane protein [17] whereas altMiD51 is located at the mitochondrial matrix [24] and both are involved in mitochondrial fission [17, 24]. AltMiD51 has also been reported to be a new assembly factor of the mitochondrial ribosome and implicated in its biogenesis [25].

Here, we employ a targeted proteomics approach based on AQUA peptides to reliably quantify the absolute amount of MiD51 and altMiD51 in two human cell lines and one human tissue, and thus we establish an improved map of the translational output of *MIEF1/SMCR7L/MiD51* by directly measuring the final protein products.

2 Experimental Procedures

2.1 Experimental design and statistical rationale

In this study, our aim is to determine the stoichiometry of two distinct proteins encoded within the gene MIEF1, the canonical protein MiD51 and altMiD51. AltMiD51 is a 70 amino acids protein coded by a short ORF in the 5'UTR region. We expect to unveil the stoichiometry of the two proteins using targeted proteomics and stable isotope labelled synthetic peptides. To determine proteotypic peptides, we first performed AP-MS DDA experiments on both proteins. Peptides were then validated in overexpressed and endogenous conditions using targeted proteomics. Selected peptides were tested for coefficient of variation (CV) using technical replicates and were validated if their CV fell below 20 %. One stable isotope peptide for each protein was purchased and tested

for linearity in a peptide mixture using two technical replicates at various concentrations. Both peptides showed a linearity range between 40 amol and 250 fmol. Absolute quantification experiments were then conducted in two cell lines (three biological replicates each) and colon tissue (three technical replicates), in MiD51 knockout HeLa cells and altMiD51 knockout HeLa cells (three biological replicates each). Protein amounts were compared using Welch's two samples t-tests.

2.2 Tissue collection and ethics

Normal human colon tissue sample was obtained from the Biobanque des maladies digestive du CRCHUS. Patient gave informed consent for the banking and use of tissue sample. The ethic review board at the Centre intégré universitaire de santé et de services sociaux de l'Estrie - Centre hospitalier universitaire de Sherbrooke (CIUSSS de l'Estrie – CHUS) approved the use of this sample for this study. Briefly, following resection for colorectal cancer, the colon was washed thoroughly and normal tissue sampling was performed at more than 10 cm of the tumour, within a region confirmed by a pathologist to be uninvolved by tumour cells (H&E). Fresh sample was flash frozen in liquid nitrogen within 30 min of surgical resection.

2.3 Cell culture

Cells were grown in Dulbecco's Modified Eagle Medium (DMEM, Wisent) supplemented with 10 % fetal bovine serum (FBS, Wisent) and antibiotic-antimycotic cocktail (Wisent). Cells were mycoplasma free (routinely tested). For transfections, cells were grown in 100 mm petri dishes until 80 % confluent and were transfected by adding 10 μg of plasmidic DNA in 2 mL of FBS/antibiotics-free DMEM and 10 μL of GeneCellIn (Eurobio) and let to grow for 24 h before cell lysis. For parallel reaction monitoring (PRM) experiments, cells were grown in 60 mm petri dishes until about 80 % confluent.

2.4 DNA constructs

DNA constructs were generated by Gibson assembly [26] of synthetic DNA (Gblocks, IDT) using the NEBuilder HiFi DNA Assembly Cloning Kit (New England BioLabs) according to manufacturer's recommendation. DNA blocks of C-terminally LAP-tagged [27] MiD51 and GFP-tagged altMiD51 were inserted separately into pcDNA 3.1(-) expression vector (Invitrogen). The context construct was built on the assembly of the full 5' region containing the altMiD51 coding sequence with a C-terminal 2 FLAG tag and the canonical MiD51 coding sequence with a C-terminal HA tag (transcript NM_019008.4) into pcDNA 3.1(-) expression vector. DNA sequences were controlled by sequencing.

2.5 Mitochondrial extracts

For western blot analysis of HeLa and CRISPR-Cas9 HeLa clones as well as PRM optimizations, mitochondrial extracts were performed according to [28] with minor modifications. Cells were grown into three 100 mm dishes until 80 % confluent, rinsed twice with PBS 1X and collected using a cell scrapper. Cells were pelleted by centrifugation at 500 g for 10 min at 4 $^\circ$C. Supernatant was discarded and cells were suspended in mitochondrial

buffer (mito-buffer : 210 mM mannitol, 70 mM sucrose, 1 mM EDTA, 10 mM HEPES-NaOH, pH 7.5, 2 mg/ml Bovine Serum Albumin (BSA), 0.5 mM PMSF and Roche EDTA-free protease inhibitor) and disrupted by passage through a 25G1 0.5 × 25 needle syringe 15 consecutive times on ice followed by a 3 min centrifugation at 2,000 g at 4 ◦ C. Supernatant was collected and the pellet was resuspendended in mitochondrial buffer. The breakage procedure was repeated four times. All four supernatant containing mitochondria were again passed through syringe needle in mito-buffer and cleared by centrifugation for 3 min at 2,000 g at 4 ◦ C. Supernatants were collected and centrifuged for 10 min at 13,000 g at 4 ◦ C to pellet mitochondrias. Pellets were washed twice with BSA-free mitochondrial buffer and pooled. Final mitochondrial pellet was lysed in SDS buffer (4% SDS, Tris-HCl 100 mM pH 7.6). After sonication, protein content was assessed using BCA assay (Pierce).

2.6 Mass spectrometry sample preparation

2.6.1 Preliminary affinity-purification (AP)

Cells were rinsed twice with cold PBS 1X and lysed with 1 mL of AP-buffer (NP-40 0.5 %, Tris-HCl 50 mM pH 7.5, NaCl 150 mM, EDTA-free Roche protease inhibitor 1X). Lysate was cleared by centrifugation (2000 g, 5 min) and supernatant was collected. GFP-Trap agarose beads (ChromoTek) were conditioned with three consecutive PBS 1X washes followed by three AP-buffer washes. Lysate supernatant was mixed with beads and incubated at 4 ◦ C for 18 h on a rotating device. Beads were then washed 3 times with AP-buffer and 5 times with 50 mM NH_4HCO_3 (ABC). Digestion was performed on beads by adding 1 μg of trypsin (Promega) in 100 μL ABC at 37 ◦ C overnight. Digestion was quenched with formic acid to a final concentration of 1 % and supernatant, containing peptides, was collected. Beads were then washed once with acetonitrile/water/formic acid (1/1/0.01 v/v) and pooled with supernatant. Peptides were dried using a speedvac, desalted using a C18 Zip-Tip (Merck) and resuspended into 25 μL of 1 % formic acid in water prior to MS analysis.

2.6.2 PRM experiments

For mitochondrial extracts, mitochondrial pellet was lysed using SDS buffer as described above. For whole cell lysates, cells were rinsed twice with cold PBS 1X and lysed using SDS buffer. Tissue sample were homogeneized using a TissueRuptor (Qiagen) in SDS buffer. Lysates were sonicated to reduce viscosity followed by a 5 min centrifugation at 14,000 g to discard debris and insoluble parts. Protein content was assessed using BCA protein assay (Pierce). A total of 100 μg of protein and 1 μg of recombinant Glutathione S-transferase (GST, Schistosoma japonicum) were reduced by adding dithiothreitol to a final concentration of 50 mM and incubated 15 min at 55 ◦ C. Lysates were prepared according to the filter aided sample preparation protocol (FASP) with minor modifications [29]. Lysates were diluted with 500 μL of 8 M urea solution and transferred into a 3 kDa centrifugation device (Amicon Ultra, Merck) and centrifuged for 30 min at 14,000 g. After one 8 M urea wash and centrifugation, samples were diluted with 200 μL of 50 mM iodoacetamide in 8 M urea and left at room temperature in the dark for 30 min. Samples were centrifuged and washed 3 times with 8 M urea. Buffer was then exchanged for 50 mM ABC with three consecutive 200 μL washes. The final retentate was digested overnight at 37 ◦ C with 1 μg of trypsin (Gold, Promega) in 40 μL ABC and AQUA [30] peptides (pepoTec Ultimate, Thermo). Tryptic peptides were collected by filter centrifugation followed by three ABC washes and centrifugation. Peptide-

containing filtrate was concentrated using a speedvac and then acidified by formic acid to a final concentration of 1 %. Peptides were desalted using a C18 Zip-Tip and dried using a speedvac.

2.6.3 Calf intestinal phosphatase treatment

Peptides were dephosphorylated using calf intestinal phosphatase (CIP) according to [31]. Briefly, 5 μg of desalted peptides were solubilized with 10 units of CIP (New England Biolabs) in 50 μL of CIP buffer (100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl2 and 1 mM DTT; pH 7.9) and incubated at 37 $^\circ$ C for 2 h. The mixture was acidified by adding trifluoroacetic acid (TFA) to a final concentration of 0.5 %. Peptides were desalted using a C18 Zip-Tip (Merck), dried and solubilized with 25 μL of 1 % formic acid in water.

2.7 nanoLC-MS/MS analysis

2.7.1 Instrument setup

A total of 12 μL of peptide mixture was loaded onto a trap column (Acclaim PepMap100 C18 column, 0.3 mm id × 50 mm, Thermo Scientific) at a constant flow rate of 4 μL/min. Peptides were separated in a PepMap C18 nano column (75 μm × 50 cm, Thermo Scientific) using a 0 – 35 % gradient (0-215 min) of 90 % acetonitrile, 0.1 % formic acid at a flow rate of 200 nL/min followed by acetonitrile wash and column re-equilibration for a total gradient duration of 4 h with a RSLC Ultimate 3000 (Dionex). Peptides were sprayed using an EASYSpray source (Thermo) at 2 kV coupled to a quadrupole-Orbitrap (QExactive, Thermo) mass spectrometer.

2.7.2 Affinity purification and MS analysis (AP-MS)

For preliminary AP-MS of GFP-tagged constructs, the mass spectrometer was used in data dependent acquisition mode (DDA). Full-MS spectra within a m/z 350-1600 mass range at 70,000 resolution were acquired with an automatic gain control (AGC) target of 1e6 and a maximum accumulation time (maximum IT) of 20 ms. Fragmentation (MS/MS) of the top ten ions detected in the Full-MS scan at 17,500 resolution, AGC target of 5e5, a maximum IT of 60 ms with a fixed first mass of 50 within a 3 m/z isolation window at a normalized collision energy (NCE) of
25. Dynamic exclusion was set to 40 s.

2.7.3 Data-dependent protein identification of AP-MS samples

Mass spectrometry RAW files were searched with Andromeda [32], search engine implemented in MaxQuant 1.5.5.1 [33]. Trypsin/P was set as digestion mode with a maximum of two missed cleavages per peptides. Oxidation of methionine and acetylation of N-terminal were set as variable modifications. Carbamidomethylation of cysteine was set as fixed modification. Precursor and fragment tolerances were set at 4.5 and 20 ppm respectively (defaults settings). Files were searched using a target-decoy approach [34] against Uniprot-Human 03/2017 release [35] and GST (92,949 entries) at a 1 % false discovery rate at peptide-spectrum-match, peptide and protein levels. Peptides sequences were recovered from MaxQuant output files.

2.7.4 PRM method refinement

Developed and applied PRM analyses were Tier 2 assays. A first PRM method was determined with a large number of peptides in order to discriminate peptides that were detectable in overexpression as well as endogenous conditions in mitochondrial extracts. Peptide unicity was first checked using neXtProt peptide uniqueness checker [36]. MS/MS spectra were then manually inspected and peptides with highest MS intensities, absence of miscleavage and high identification scores were selected for preliminary PRM peptide evaluation. The peptide list consisted in 30 mass over charges corresponding to unique peptides of MiD51 (11 peptides), altMiD51 (4 peptides), HSP60 (4 peptides) and GST (5 peptides) at various charge states. Method consisted in a Full-MS spectra acquisition with an AGC target of 3e6, maximum IT of 70 ms and a resolution of 70,000 followed by an unscheduled targeted-MS2 method with an AGC target of 5e5 ions, maximum IT of 130 ms, resolution of 17,500 with a 2 m/z isolation window and a NCE of 27.

A second method was used to evaluate the signal recovered after CIP treatment on endogenous mitochondrial extracts. The peptide list consisted in 11 mass over charges based on previous PRM experiments corresponding to peptides of MiD51 (3 peptides), altMiD51 (2 peptides), HSP60 (3 peptides) and GST (3 peptides). Method consisted in a Full-MS spectra acquisition with an AGC target of 3e6, maximum IT of 70 ms and a resolution of 70,000 followed by an unscheduled targeted-MS2 method with an AGC target of 5e5 ions, maximum IT of 150 ms, resolution of 17,500 with a 2 m/z isolation window and a NCE of 27. All method optimization files were processed using Skyline [37].

2.7.5 High sensitivity PRM

For endogenous CV analysis in whole cell extracts and mitochondrial extracts as well as absolute quantification experiments, mass spectrometer was set for highest sensitivity according to [38]. Method consisted into a Full-MS spectra acquisition with an AGC target of 3e6, maximum IT of 70 ms and a resolution of 70,000 followed by an unscheduled targeted MS2 method with an AGC target of 1e6 ions, maximum IT of 250 ms resolution of 70,000 and a NCE of 27. Isolation list contained one peptide from altMiD51, one for MiD51 and their AQUA standards, one peptide from GST spike-in as well as one peptide from HSP60 which were used as sample processing controls.

2.7.6 High sensitivity PRM sample analysis

Mass-spectrometry RAW files were analyzed using Xcalibur 2.2 (Thermo) by measuring area of each peptide monoisotopic transitions within a 3 ppm mass precision window. For AQUA peptide calibration curves, internal standards were spiked into a HeLa digest and analyzed with high sensitivity PRM in conditions described above. For each peptide, five precursor to fragment transitions starting from N-terminus within a mass deviation of 3 ppm were assessed for linearity and CV analysis, considering that a transition with a CV below 20 % at a given concentration is quantifiable. For endogenous CV analysis, most quantifiable precursor to fragment transitions were measured for each peptide within a 3 ppm precision window, and two replicates were compared. For absolute quantification experiments, protein concentration was determined by comparing the ratio of the endogenous peptide to spiked-in AQUA standard and its concentration with the same precursor to fragment transitions within a 3 ppm mass precision window. Peptide ratios were kept below 25. Spectral similarity was

controlled by importing RAW files into Skyline and peptides were validated if their spectral contrast angles [39] or ratio dot products were close to 1 as well as their retention times matching AQUA standards.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [40] partner repository with the dataset identifier PXD008147 (User : reviewer60166@ebi.ac.uk, Password : rOUnIeS9).

2.8 CRISPR-Cas9-mediated MiD51 and altMiD51 knockout (KO)

2.8.1 Knockouts clonal cell generation

CRISPR-Cas9-mediated MiD51 and altMiD51 KO HeLa cells were generated according to [41] with minor modifications. Briefly, sgRNAs were designed using the Broad Institute sgRNA Designer (CRISPRko) tool (http://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design, [42]) and confirmed with the CCTOP tool (http://crispr.cos.uni-heidelberg.de/, [43]). CRISPR-Cas9 related oligonucleotides are described in Table 1. The sgRNA inserts, containing an extra G in 5' required for the U6 RNA polymerase III promoter, were prepared by annealing the top and bottom oligos (Table 1) and cloned into the pSpCas9(BB)-2A-GFP plasmid (Addgene #48138, [41]). The resulting plasmids were verified by sequencing. Enrichment for Cas9-2A-GFP expressing cells and isolation of clonal cell populations were performed 24 h after transfection by single-cell FACS sorting. The initial validation of genome editing was done by Mismatch-cleavage assay using T7 Endonuclease I (NEB), GenElute Mammalian Genomic DNA Miniprep Kit (Sigma) with missmatch assays primers (Table 1). Cells edition was confirmed by western blotting (Figure 3) and by sequencing PCR amplicons derived from the target sites.

2.8.2 Characterization of heterozygote MiD51 KOs

Genomic DNA was amplified using MiD51 mismatch assays primers with primer extension allowing its insertion into linearized (EcoRI, BamHI, New England Biolabs) pcDNA 3.1(-) expression vector *via* Gibson assembly as mentioned above. Plasmids were purified and sequenced.

2.9 Western blotting

For each sample, 50 μg of mitochondrial protein extract was mixed 1/1 (v/v) with Laemmli buffer (4 % SDS w/v, 20 % glycerol v/v, Tris-HCl 100 mM pH 6.8, 5 % β-mercapto ethanol v/v) and heated at 95 $^\circ$C for 15 min. For altMiD51, proteins were separated in a 4 % stacking/15 % acrylamide-bisacrylamide (29/1 w/w) resolving SDSPAGE for one hour at 200 V constant voltage using a glycine-buffer. For MiD51, proteins were separated in a 4 % stacking/10 % acrylamide-bisacrylamide (49.5 % T, 3 % C) resolving tricine SDS-PAGE [15, 44] gel (16×18 cm) for 18 h using 0.2 M Tris-HCl pH 8.9 as anode buffer and 0.1 M Tris-HCl, 0.1 M tricine, 0.1 % SDS pH 8.25 as cathode buffer (25 mA constant current). Proteins were transferred onto polyvinyldiene difluoride membranes. The membranes were blocked wita 5 % milk supplemented Tris-buffered saline 0.2 % Tween-20 (TBST). Membranes were probed with a custom anti-altMiD51 rabbit antibody (Proteintech, see below), a polyclonal anti MiD51 rabbit antibody (Proteintech 20164-1-AP) and a mouse monoclonal anti-mitochondrial HSP 70 antibody (MA3028, Pierce) at 4 $^\circ$C overnight. The membrane was then washed three times with TBST and probed with

goat anti-mouse (sc-2005, Santa-Cruz Biotechnology) or goat anti-rabbit-conjugated horseradish peroxidase antibodies (7074S, Cell Signaling Technology).

2.10 AltMiD51 antibodies

Rabbit polyclonal anti-altMiD51 antibodies were raised against the full-length 70 amino acids recombinant altMiD51 protein and affinity purified (Proteintech Group Inc.).

2.11 Statistics

All graphics and statistics were made using R [45] 3.3.2 and ggplot2 [46] 2.2.1 or higher.

2.12 Cross validation with elongating ribosomes

Ribo-Seq coverage of both ORFs were extracted from GWIPS [47] for HEK and HeLa cells. All nucleotides of both ORFs were considered as mappable using Umap track of UCSC Genome Browser [48] and ribosome densities were compared between altMiD51 and MiD51 ORFs.

3 Results

3.1 Determination of MiD51 and altMiD51 proteotypic peptides

In addition to the canonical CDS (Consensus CDS CCDS13995; RefSeq NM_019008.5; Ensembl ENST00000325301) and associated protein (RefSeq NP_061881, UniProt Q9NQG6; Ensembl ENSP00000327124), human *MIEF1* contains a functional and recently annotated altORF (GenBank HF548110) coding for a small protein (Uniprot L0R8F8; GenBank CCO13821.1; Ensembl ENSP00000490747) (Figure 1 a & b). Thus, *MIEF1* is clearly a prototypical dual-coding gene for which the absolute quantification of the large and small protein products is unknown. We evaluated the ability of MiD51 and altMiD51 to generate proteotypic peptides after trypsin digestion. Proteotypic peptides are specific for each protein and they must be consistently detected with excellent quality precursor and fragment mass transitions [49–52]. In order to facilitate the detection of specific tryptic peptides for both proteins, we used affinity purification coupled with mass spectrometry. MiD51^{GFP} and altMiD51^{GFP} were independently overexpressed in HeLa cells. Both proteins were affinity purified and analyzed via data-dependant (DDA) nano capillary liquid chromatography mass spectrometry (nanoLC-MS/MS). Several proteotypic peptides were detected for a total sequence coverage of 68.6 % and 71.9 % for altMiD51 and MiD51, respectively (Figure 1 b & Supplementary data 1). After manual evaluation (see method section), best quality proteotypic peptides were selected for parallel reaction monitoring (PRM) optimization [38, 53].

3.2 PRM optimization for MiD51 and altMiD51 proteotypic peptides

Selected proteotypic peptides were then validated in low-sensitivity PRM experiments in 3 kDa FASP processed samples [29]. Indeed, altMiD51 is a small protein of 70 amino acids and a low M.W. cut-off is necessary to ensure protein retention during sample preparation. Since both MiD51 and altMiD51 are mitochondrial proteins [24],

mitochondria were isolated from mock-transfected cells and from cells transfected with a cDNA containing both the CDS coding for MiD51 and the native 5'UTR containing the altORF coding altMiD51 (RefSeq transcript NM_019008.4). Two proteotypic peptides for each protein were detected in these mitochondrial extracts (Supplementary Figure 1). Signal intensity for endogenous mitochondrial HSP60 peptide shows that the protein concentration of mock and transfected mitochondrial extracts were similar, and that the intensity difference for altMiD51 and MiD51 peptides between mock-transfected and altMiD51/MiD51-transfected samples did not result from differences in mitochondria preparation.

As MiD51's most intensely detected peptide (AISAPTSPTR) bore known phosphosites (Figure 1 b), a second PRM method including a dephosphorylation step using calf intestinal phosphatase (CIP) was implemented [31]. A fraction of MiD51 was indeed phosphorylated since CIP treatment resulted in a 20 % increase in intensity for AISAPTSPTR (Supplementary Figure 2). The efficiency of CIP treatment was validated with two known HSP60 tryptic phosphorylated peptides, VGGTSDVEVNEK and VTDALNATR (phosphosite.org) with an increase in intensity of 103 % and 133 %, respectively. The intensity of a non-phosphorylated HSP60 peptide did not change significantly. AltMiD51 peptides are clearly non phosphorylated since CIP treatment did not change significantly their intensity (Supplementary Figure 2). Based on these results, we selected peptides EAVLSLYR and AISAPTSPTR for absolute quantitation of altMiD51 and MiD51, respectively.

Finally, the precision of the most sensitive PRM method across different samples was estimated with the measure of the coefficient of variation (CV) on mitochondrial and whole cell extracts. Indeed, a CV below 20 % is required for absolute quantification [54]. The CVs were systematically below 20 %, indicating that both mitochondrial and whole cell extracts were suitable for quantification (Supplementary Figure 3). Even though peptide intensities are higher in mitochondrial extracts, we decided to use whole cell lysates for absolute quantification of altMiD51 and MiD51 as their preparation does not involve cell fractionation, with the risk of variable mitochondrial recovery.

3.3 AltMiD51 and MiD51 protein abundances

Two synthetic stable isotope-labeled peptides for absolute quantification (AQUA) [30], EAVLSLYR and AISAPTSPTR, were spiked into the protein sample after trypsin digestion from HeLa cells and analyzed *via* PRM (Figure 1 b). A total of 5 y ion transitions starting from the most N-terminal amino acid were measured and both peptides displayed at least one quantifiable transition within a range of 40 amol - 250 fmol and a CV < 20 % (Supplementary Figure 4).

Absolute quantification PRM experiments were performed by spiking AQUA peptides with trypsin into the digestion mixture as described by [30]. After desalting and dephosphorylation with CIP treatment, the resulting peptides were processed using a high sensitivity PRM method [38]. For each peptide, retention times for the corresponding native and AQUA species as well as spectral contrast angles or ratio dot product [39] were controlled to ensure correct identification (Figure 2 a, b & Supplementary Figure 6-12). The absolute amount of native peptides were thus determined (Supplementary data 2).

3.4 CRISPR-Cas9-mediated independent inactivation of altMiD51 or MiD51

As this is the first absolute quantification of a large and small protein encoded by two independent ORFs in the same gene, it is important to show that absolute amounts of MiD51 and altMiD51 are partially or completely obliterated by inactivating their respective coding sequences. Experimental modulation of altMiD51 expression independently of MiD51 expression using a RNAi-based knockdown approach is impossible since both proteins are coded by the same gene, and both coding sequences are present in the same transcripts. This is a general challenge for the study of small and large proteins coded in the same gene [14]. Thus, we implemented a CRISPRCas9 approach to independently prevent the expression of either altMiD51 or MiD51 (Figure 3 a & b) [41, 55].

Genome-edited clonal cell lines were validated by sequencing the targeted genomic region. The sequence of the PCR-amplified altMiD51 genomic region confirmed the homozygous 1 bp insertion of a A/T at position 40 of exon 2, at the Cas9 cleavage site (Figure 3 c). For MiD51, the sequence electropherogram of the PCR-amplified genomic region showed overlapping peaks (Figure 3 c), indicating the presence of heterozygous mutations in the different alleles.

AltMiD51 was competley undetectable both by western blot (Figure 3 d) and absolute quantification (Figure 4 a, Welch's t-test p-value = 0.0013), confirming successful editing of the altMiD51 ORF (Figure 3 c). Remarkably, levels of MiD51 were significantly increased in altMiD51-edited cells (Figure 4 a, Welch's t-test p-value = 0.0006). Although MiD51 was not detected by western blot in CRISPR-Cas9 MiD51 edited cells (Figure 3 d), PRM analyses showed a 86 % reduction in MiD51 levels (Figure 4 a, Welch's t-test p-value = 0.0004), suggesting that non-edited WT alleles remained. However, sequencing alleles of MiD51-edited HeLa (Figure 3 e) revealed that no WT sequence was detected, suggesting that signal from PRM experiments is due to the 6 nucleotides, and thus 2 amino acids, loss in MiD51 sequence, giving a non-frameshifted sequence coding for a truncated protein containing the AQUA MiD51 peptide (Figure 3 e, blue bar). Overall, genome editing of altMiD51 and MiD51 conclusively validated the proteotypic peptides selected for absolute quantification, and the presence of two functional and physically independent coding informations in the same gene.

3.5 Absolute amounts and ratio of altMiD51 to MiD51

Absolute quantification performed in HEK 293, HeLa and colon tissue indicate that altMiD51 is the most abundant protein product of *MIEF1* (Figure 4 a, Supplementary data 2). We compared absolute quantities of altMiD51 and MiD51 in HEK 293, HeLa and human colon tissue samples to determine their stoechiometric relationship. The stoichiometry indicated that the most abundant translation product from *MIEF1* is altMiD51 rather than the canonical MiD51 protein. The ratio of altMiD51 to MiD51 is 2.71 in HEK 293 cells, 5.73 in HeLa cells, and 2.62 in Human colon tissue (Figure 4 b & Supplementary Figure 13). This observation is consistent with our analysis of ribosome occupancy with data extracted from GWIPS (Supplementary Figure 14; Supplementary data 3).

4 Discussion

Ribosome profiling and proteogenomics strongly support the translation of alternative protein products from altORFs in addition to the translation of canonical CDSs. Yet, the absolute quantification of a small and a large protein coded by the same gene is unknown. Here, we show that levels of the 70 amino acid altMiD51, a small protein encoded in an exon originally annotated as "non-coding" of *MIEF1/SMCR7L/MiD51* are two to six times higher than the levels of the canonical MiD51 protein in cells and in a human tissue. This work illustrates that small proteins are important contributors of the proteome, and it is not because that altORFs and alternative proteins are not annotated, unlike large proteins, that they do not exist. Obviously, it is very likely that this is not a general feature of altORFs and that the expression levels of small and large proteins coded by the same genes are highly variable and gene-specific. Also, there is no correlation between protein abundance and functionality, and because the ratio altMiD51/MiD51 is > 2 does not mean that the function of altMiD51 is more significant than that of MiD51.

Several physiological processes could explain the higher ratio of altMiD51 to MiD51, including a difference in protein synthesis, a difference in protein degradation or a combination of both. Nonetheless, according to the scanning model for translation initiation, the most likely mechanism is the localization of altMiD51 upstream of MiD51 that would favor altMiD51 translation. This hypothesis is supported by ribosome profiling data aligned to the *MIEF1* locus which indicate that the density of elongating ribosomes is higher on the altORF compared to the CDS [21, 23, 47], suggesting that ribosomes efficiently translate altMiD51. In addition, MIEF1 is moderately resistant to eIF2 repression in response to severe stress induced by sodium arsenite [21]. Genes resistant to eIF2 repression are characterized by the presence of an efficiently translated upstream ORF and partial repression of translation of the main CDS in normal conditions, and derepression in response to environmental stresses [56]. Our proteomics data are in agreement with the dampening of MiD51 translation under physiological conditions.

CRISPR-mediated altMiD51 and MiD51 KO experiments resulted in two important observations. First, we observed that MiD51 expression was significantly increased in altMiD51 KO cells (Figure 4 a). In these cells, the single bp (A/T) insertion in Cas9-edited altMiD51 coding sequence resulted in the truncation of altMiD51 ORF from 210 bps to 78 bps and a parallel increase of intercistronic distance altMiD51-MiD51 from 98 to 231. The combination of a shorter upstream ORF and a longer intercistronic distance were previously shown to increase re-initiation of the downstream ORF [12, 57–60]. Thus, in addition to its role as a coding sequence for a novel mitochondrial fission factor [24] and an assembly factor implicated in mitoribosomal biogenesis [25], altMiD51 ORF may function as an upstream ORF regulating the translation of MiD51. Second, MiD51 KO cells still express altMiD51 at normal levels, which demonstrates that knocking out the canonical CDS does not completely inactivate *MIEF1*. This result illustrates for the first time that inactivating an annotated CDS may not necessarily obliterate a gene.

A combination of several circumstances have allowed small proteins to go unnoticed until recently. First, according to current human annotations, protein-coding genes have a single CDS, generally the longest ORF [1]. Thus, all efforts to find the physiological function or role in the pathology of a specific gene are invariably focused

on the protein encoded by this CDS, or one of its variants generated by alternative splicing. Second, in the absence of annotation of non-canonical ORFs, the protein sequence of the corresponding proteins cannot be routinely detected by MS-based proteomics approaches which rely on current protein databases containing the sequences of canonical proteins only. Third, the widely used western blot technique relies on specific antibodies, but antibodies have been raised and commercialized for canonical proteins only. Raising novel specific antibodies may take time and several attempts, thus delaying the investigations on small proteins. Fourth, the detection of small proteins by MS-based proteomics is more challenging than for large proteins. Typically, the proteome has to be fractionnated to enrich low molecular weight proteins, and the identification often relies on a single tryptic peptide [5, 9]. In addition, there may be no sites for trypsin digestion and peptides exceeding 25 aa are rarely identified in bottom-up proteomics. Fifth, because they are short, small proteins are less likely to have known protein domains discovered in large proteins, or to display a specific structure. Thus, there might exist a biased perception that small proteins have minor functions compared to large proteins in biological mechanisms. Yet, many small proteins have essential functions in prokaryotes and eukaryotes [14, 61].

AltMiD51 was integrated into the automatically annotated UniProtKB/TrEMBL database (identifier L0R8F8) in March 2013, following its detection under the name altSMCR7L [3]. It was integrated into the manually annotated UniProtKB/Swiss-Prot database in March 2017. Similar to altMiD51, which is now a manually annotated bicistronic gene, it will be important to update genome annotations according to recent proteogenomics studies [14, 24]. Indeed, the function of a dual-coding gene should not be inferred according to the molecular activity of the larger protein product only. In addition, the impact of mutations on gene function should not be analyzed in the conceptual frame of a single CDS, since mutations outside currently annotated CDSs may affect non-canonical ORFs and ultimately, gene function [62]. Finally, our results clearly demonstrate that knocking out the canonical CDS in a gene and leaving altORFs unnaltered does not completely abrogate the translation output of that gene.

5 Acknowledgments

This study was supported by the Biobanque des maladies digestives du Centre de recherche du CHUS (CIUSSS de l'Estrie – CHUS), certified by the Canadian Tissue Repository Network and affiliated with the Réseau de recherche sur le cancer. Authors are thankful to Michael T. Ryan and Laura Osellame for constructive exchanges, particularly on MiD51 detection via western-blot, François-Michel Boisvert for access to mass spectrometer. Recombinant GST was a generous gift of Marie-Line Dubois.

References

[1] Marcel E Dinger, Ken C Pang, Tim R Mercer, and John S Mattick. Differentiating protein-coding and noncoding rna: challenges and ambiguities. *PLoS computational biology*, 4(11):e1000176, 2008.

[2] Gloria A Brar and Jonathan S Weissman. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology*, 16(11):651–664, 2015.

[3] Benoît Vanderperre, Jean-François Lucier, Cyntia Bissonnette, Julie Motard, Guillaume Tremblay, Solène Vanderperre, Maxence Wisztorski, Michel Salzet, François-Michel Boisvert, and Xavier Roucou. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS one*, 8(8):e70698, 2013.

[4] Gerben Menschaert, Wim Van Criekinge, Tineke Notelaers, Alexander Koch, Jeroen Crappé, Kris Gevaert, and Petra Van Damme. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & cellular proteomics*, 12(7):1780–1790, 2013.

[5] Jiao Ma, Carl C Ward, Irwin Jungreis, Sarah A Slavoff, Adam G Schwaid, John Neveu, Bogdan A Budnik, Manolis Kellis, and Alan Saghatelian. Discovery of human sorf-encoded polypeptides (seps) in cell lines and tissue. *Journal of proteome research*, 13(3):1757–1765, 2014.

[6] Alexander Koch, Daria Gawron, Sandra Steyaert, Elvis Ndah, Jeroen Crappé, Sarah De Keulenaer, Ellen De Meester, Ming Ma, Ben Shen, Kris Gevaert, et al. A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*, 14(23-24):2688–2698, 2014.

[7] Ariel A Bazzini, Timothy G Johnstone, Romain Christiano, Sebastian D Mackowiak, Benedikt Obermayer, Elizabeth S Fleming, Charles E Vejnar, Miler T Lee, Nikolaus Rajewsky, Tobias C Walther, et al. Identification of small orfs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal*, page e201488411, 2014.

[8] Alexey I Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature methods*, 11(11):1114–1125, 2014.

[9] Jiao Ma, Jolene K Diedrich, Irwin Jungreis, Cynthia Donaldson, Joan Vaughan, Manolis Kellis, John R Yates III, and Alan Saghatelian. Improved identification and analysis of small open reading frame encoded polypeptides. *Analytical chemistry*, 88(7):3967–3975, 2016.

[10] Volodimir Olexiouk and Gerben Menschaert. Identification of small novel coding sequences, a proteogenomics endeavor. In *Proteogenomics*, pages 49–64. Springer, 2016.

[11] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223, 2009.

[12] Hélène Mouilleron, Vivian Delcourt, and Xavier Roucou. Death of a dogma: eukaryotic mrnas can code for more than one protein. *Nucleic acids research*, 44(1):14–23, 2015.

[13] Nicholas T Ingolia. Ribosome footprint profiling of translation throughout the genome. *Cell*, 165(1):22–33, 2016.

[14] Vivian Delcourt, Antanas Staskevicius, Michel Salzet, Isabelle Fournier, and Xavier Roucou. Small proteins encoded by unannotated orfs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mrna. *Proteomics*, 2017.

[15] Catherine S Palmer, Laura D Osellame, David Laine, Olga S Koutsopoulos, Ann E Frazier, and Michael T Ryan. Mid49 and mid51, new components of the mitochondrial fission machinery. *EMBO reports*, 12(6): 565–573, 2011.

[16] Zhenzhen Zhang, Lei Liu, Shengnan Wu, and Da Xing. Drp1, mff, fis1, and mid51 are coordinated to mediate mitochondrial fission during uv irradiation–induced apoptosis. *The FASEB Journal*, 30(1):466–476, 2016.

[17] Laura D Osellame, Abeer P Singh, David A Stroud, Catherine S Palmer, Diana Stojanovski, Rajesh Ramachandran, and Michael T Ryan. Cooperative and independent roles of the drp1 adaptors mff, mid49 and mid51 in mitochondrial fission. *J Cell Sci*, 129(11):2170–2181, 2016.

[18] Sooncheol Lee, Botao Liu, Soohyun Lee, Sheng-Xiong Huang, Ben Shen, and Shu-Bing Qian. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 109(37):E2424–E2432, 2012.

[19] Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, Dhanashree S Kelkar, Ruth Isserlin, Shobhit Jain, et al. A draft map of the human proteome. *Nature*, 509(7502):575–581, 2014.

[20] Jeroen Crappé, Elvis Ndah, Alexander Koch, Sandra Steyaert, Daria Gawron, Sarah De Keulenaer, Ellen De Meester, Tim De Meyer, Wim Van Criekinge, Petra Van Damme, et al. Proteoformer: deep proteome coverage through ribosome profiling and ms integration. *Nucleic acids research*, 43(5):e29–e29, 2014.

[21] Dmitry E Andreev, Patrick BF O'Connor, Ciara Fahey, Elaine M Kenny, Ilya M Terenin, Sergey E Dmitriev, Paul Cormican, Derek W Morris, Ivan N Shatsky, and Pavel V Baranov. Translation of 5' leaders is pervasive in genes resistant to eif2 repression. *Elife*, 4:e03971, 2015.

[22] Carmela Sidrauski, Anna M McGeachy, Nicholas T Ingolia, and Peter Walter. The small molecule isrib reverses the effects of eif2α phosphorylation on translation and stress granule assembly. *Elife*, 4:e05033, 2015.

[23] Lorenzo Calviello, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zauber, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler. Detecting actively translated open reading frames in ribosome profiling data. *Nature methods*, 13(2):165, 2016.

[24] Sondos Samandi, Annie V Roy, Vivian Delcourt, Jean-François Lucier, Jules Gagnon, Maxime C Beaudoin, Benoît Vanderperre, Marc-André Breton, Julie Motard, Jean-François Jacques, et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, 6:e27860, 2017.

[25] Alan Brown, Sorbhi Rathore, Dari Kimanius, Shintaro Aibara, Xiao-chen Bai, Joanna Rorbach, Alexey Amunts, and V Ramakrishnan. Structures of the human mitochondrial ribosome in native states of assembly. *Nature structural & molecular biology*, 24(10):866–869, 2017.

[26] Daniel G Gibson, Lei Young, Ray-Yuan Chuang, J Craig Venter, Clyde A Hutchison, and Hamilton O Smith. Enzymatic assembly of dna molecules up to several hundred kilobases. *Nature methods*, 6(5):343–345, 2009.

[27] Iain M Cheeseman and Arshad Desai. A combined approach for the localization and tandem affinity purification of protein complexes from metazoans. *Sci. STKE*, 2005(266):pl1, 2005.

[28] Bruno Antonsson, Sylvie Montessuit, Belen Sanchez, and Jean-Claude Martinou. Bax is present as a high molecular weight oligomer/complex in the mitochondrial membrane of apoptotic cells. *Journal of Biological Chemistry*, 276(15):11615–11623, 2001.

[29] Jacek R Wisniewski, Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. Universal sample preparation method for proteome analysis. *Nature methods*, 6(5):359, 2009.

[30] Scott A Gerber, John Rush, Olaf Stemman, Marc W Kirschner, and Steven P Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proceedings of the National Academy of Sciences*, 100(12):6940–6945, 2003.

[31] Ronghu Wu, Wilhelm Haas, Noah Dephoure, Edward L Huttlin, Bo Zhai, Mathew E Sowa, and Steven P Gygi. A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nature methods*, 8 (8):677–683, 2011.

[32] Jurgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: a peptide search engine integrated into the maxquant environment. *Journal of proteome research*, 10(4):1794–1805, 2011.

[33] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.

[34] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207–214, 2007.

[35] UniProt. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, 2017.

[36] Mathieu Schaeffer, Alain Gateau, Daniel Teixeira, Pierre-André Michel, Monique Zahn-Zabal, and Lydie Lane. The nextprot peptide uniqueness checker: a tool for the proteomics community. *Bioinformatics*, 33 (21):3471–3472, 2017.

[37] Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. Skyline: an open source

document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, 2010.

[38] Sebastien Gallien, Adele Bourmaud, Sang Yoon Kim, and Bruno Domon. Technical considerations for largescale parallel reaction monitoring analysis. *Journal of proteomics*, 100:147–159, 2014.

[39] Katty X Wan, Ilan Vidavsky, and Michael L Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry*, 13(1):85–88, 2002.

[40] Juan Antonio Vizcaíno, Attila Csordas, Noemi Del-Toro, José A Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, Yasset Perez-Riverol, Florian Reisinger, Tobias Ternent, et al. 2016 update of the pride database and its related tools. *Nucleic acids research*, 44(D1):D447–D456, 2015.

[41] F Ann Ran, Patrick D Hsu, Jason Wright, Vineeta Agarwala, David A Scott, and Feng Zhang. Genome engineering using the crispr-cas9 system. *Nature protocols*, 8(11):2281–2308, 2013.

[42] John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2):184–191, 2016.

[43] Manuel Stemmer, Thomas Thumberger, Maria del Sol Keyer, Joachim Wittbrodt, and Juan L Mateo. Cctop: an intuitive, flexible and reliable crispr/cas9 target prediction tool. *PloS one*, 10(4):e0124633, 2015.

[44] Hermann Schägger. Tricine–SDS-PAGE. *Nature Protocols*, 1(1):16–22, 2006.

[45] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[46] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

[47] Audrey M Michel, Gearoid Fox, Anmol M. Kiran, Christof De Bo, Patrick BF O'Connor, Stephen M Heaphy, James PA Mullan, Claire A Donohue, Desmond G Higgins, and Pavel V Baranov. Gwips-viz: development of a ribo-seq genome browser. *Nucleic acids research*, 42(D1):D859–D864, 2013.

[48] Mehran Karimzadeh, Carl Ernst, Anshul Kundaje, and Michael M Hoffman. Umap and bismap: quantifying genome and methylome mappability. *bioRxiv*, page 095463, 2016.

[49] Bernhard Kuster, Markus Schirle, Parag Mallick, and Ruedi Aebersold. Scoring proteomes with proteotypic peptide probes. *Nature reviews Molecular cell biology*, 6(7):577–583, 2005.

[50] Parag Mallick, Markus Schirle, Sharon S Chen, Mark R Flory, Hookeun Lee, Daniel Martin, Jeffrey Ranish, Brian Raught, Robert Schmitt, Thilo Werner, et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nature biotechnology*, 25(1):125–131, 2007.

[51] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582, 2014.

[52] Daniel P Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, et al. Building proteometools based on a complete synthetic human proteome. *Nature Methods*, 2017.

[53] Adele Bourmaud, Sebastien Gallien, and Bruno Domon. Parallel reaction monitoring using quadrupoleorbitrap mass spectrometer: Principle and applications. *Proteomics*, 16(15-16):2146–2159, 2016.

[54] Sebastien Gallien, Sang Yoon Kim, and Bruno Domon. Large-scale targeted proteomics using internal standard triggered-parallel reaction monitoring (is-prm). *Molecular & Cellular Proteomics*, 14(6):1630–1644, 2015.

[55] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, 2007.

[56] Sara K Young and Ronald C Wek. Upstream open reading frames differentially regulate gene-specific translation in the integrated stress response. *Journal of Biological Chemistry*, pages jbc–R116, 2016.

[57] MARILYN Kozak. Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Molecular and Cellular Biology*, 7(10):3438–3445, 1987.

[58] BG Luukkonen, Wei Tan, and Stefan Schwartz. Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mrnas is determined by the length of the upstream open reading frame and by intercistronic distance. *Journal of Virology*, 69(7):4086–4094, 1995.

[59] Marilyn Kozak. Constraints on reinitiation of translation in mammals. *Nucleic Acids Research*, 29(24): 5226–5232, 2001.

[60] Cristina Barbosa, Isabel Peixeiro, and Luísa Romão. Gene expression regulation by upstream open reading frames and human disease. *PLoS genetics*, 9(8):e1003529, 2013.

[61] Gisela Storz, Yuri I Wolf, and Kumaran S Ramamurthi. Small proteins can no longer be ignored. *Annual review of biochemistry*, 83:753–777, 2014.

[62] Marie A Brunet, Sébastien A Levesque, Darel J Hunting, Alan A Cohen, and Xavier Roucou. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome research*, 2018.

[63] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2014.

6 Figure Legends

Figure 1: Schematic representation of human *MIEF1* RefSeq variant 1 mRNA and altMiD51 and MiD51 proteins
(a) Human *MIEF1* includes 11 exons (RefSeq, GRCh38.p7). The mRNA variant 1 (NM_019008.5) shown here contains 6 exons. The CDS (blue) is shared between exons 3 to 6. AltMiD51 ORF is localized within exon 2, annotated as a non-coding exon. (b) Sequence coverage in AP-MS experiments is represented in light colors. Proteotypic peptides sequence and positions (a.a.) are shown. EAVLSLYR and AISAPTSPTR peptides were selected for absolute quantification. * : known phosphorylated residue (phosphosite.org, [63])

Figure 2: Identification and quantification of altMiD51 and MiD51
(a) Extracted fragment-ion transition chromatograms of MiD51 ($[AISAPTSPTR+2H]^{2+} \rightarrow y_6^+$) and altMiD51 ($[EAVLSLYR+2H]^{2+} \rightarrow y_4^+$) peptides in HeLa cells. (b) Spectral contrast angle analysis of endogenous peptides (top) and stable isotope labelled synthetic peptides (bottom) extracted from Supplementary Figure 5.

Figure 3: CRISPR-Cas9 editing of genomic altMiD51 and MiD51
(a) Schematic representation of CRISPR-Cas9 experiments strategy. For clarity, only 5 exons are shown. Alt-MiD51 within exon 2 is shown in red. MiD51 coding sequence (blue), overlaps exons 3, 4 and part of exon 5. (b) Genomic sequences around the programmed cut sites in non-edited HeLa cells and corresponding sequences. PAM sites are highlighted in yellow, and the genomic sequences targeted by the guide RNAs are highlighted in green. The programmed cut sites are also shown, at nucleotide 40 in exon 2 and nucleotide 137 in exon 3. (c) Genomic sequence around the programmed cut sites in CRISPR-Cas9-edited HeLa cells. In the altMiD51-edited clone, a 1 bp A/T insertion (labeled in red, and red star above the electropherogram) occurred at the cut site. In the MiD51edited clone, a mixture of different sequences are detected 10 nucleotides upstream the programmed cut site (red star), indicating the presence of different alleles. (d) Mitochondrial extracts from non-edited, MiD51-edited and altMiD51-edited HeLa cells were lysed and analyzed by western blot with antibodies against mtHSP70, MiD51 and custom altMiD51 antibodies, as indicated. (e) CRISPR-Cas9 MiD51 knock-out sequence analysis. Sequences are aligned with electropherogram of panel c. * refers to a non-specific target of MiD51 antibodies [17]

Figure 4: Absolute quantification and stoichiometries
(a) Absolute quantification of altMiD51 and MiD51 in Colon tissue (technical triplicate), HEK 293, HeLa and CRISPR-Cas9 knock outs (biological triplicates). Error bars = standard deviations. (b) Stoichiometry determination based on absolute quantities of altMiD51 and MiD51. Boxplots represent three biological (HEK 293 & HeLa) or technical (Colon) replicates. Welch's t-test $p < 0.05$ (*), < 0.01 (**), < 0.001 (***)

7 Table

Table 1: Oligonucleotide sequences used for CRISPR-Cas9 genome editing experiments.

	Oligonucleotide sequences for altMiD51 knock out	Oligonucleotide sequences for MiD51 knock out
Genomic target site	5′-TGGAGCCGAGAGGCGGTGCT-3′	5′-CGCTGGCAGTTAAGCGGGTA-3′
Top oligonucleotide	5′-CACCGAGCACCGCCTCTCGGCTCCA-3′	5′-CACCGTACCCGCTTAACTGCCAGCG-3′
Bottom oligonucleotide	5′-AAACTGGAGCCGAGAGGCGGTGCTC-3′	5′-AAACCGCTGGCAGTTAAGCGGGTAC-3′
T7 endonuclease 1 mismatch assays primers	5′-GGGGTCTCTGGAACTTGGAT-3′ 5′-TCCTTTTCTCGGTCCCTTGC-3′	5′-GGTCCCAGTACTTATGGCCG-3′ 5′-CCACGCAGAAAATCTCAGGG-3′

8 Figures

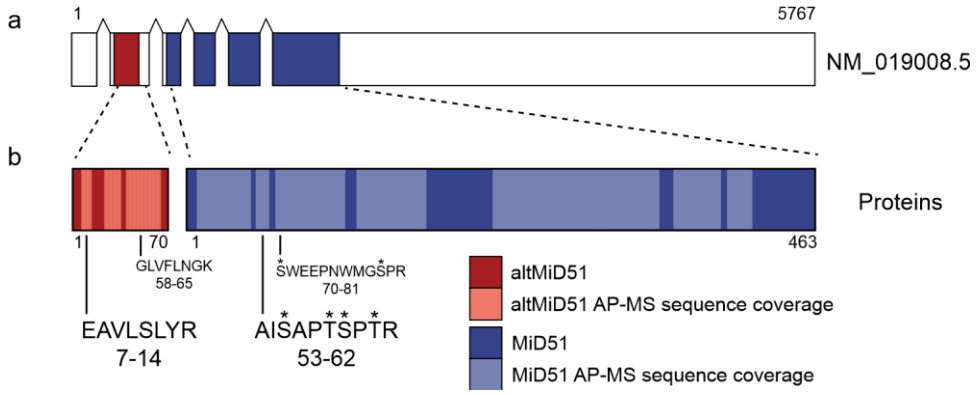

Figure 1

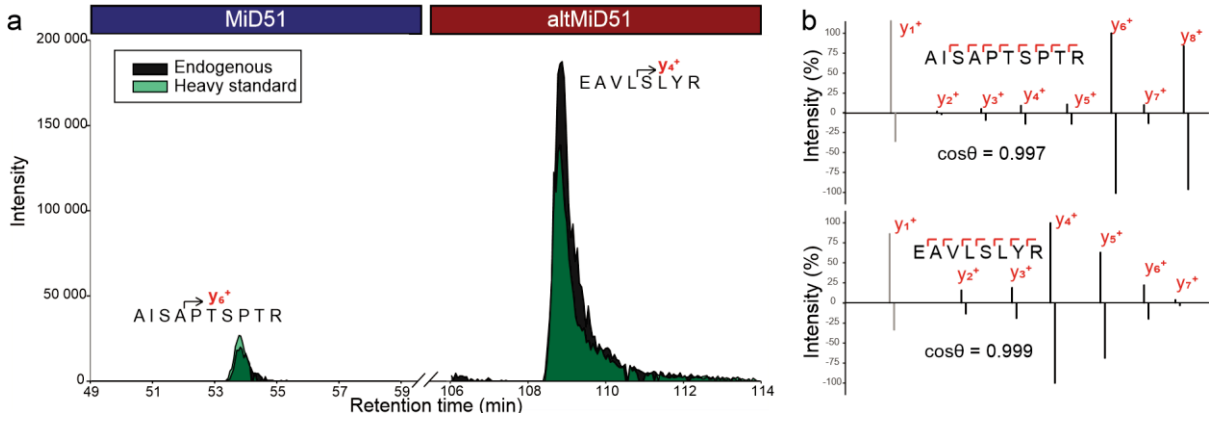

Figure 2

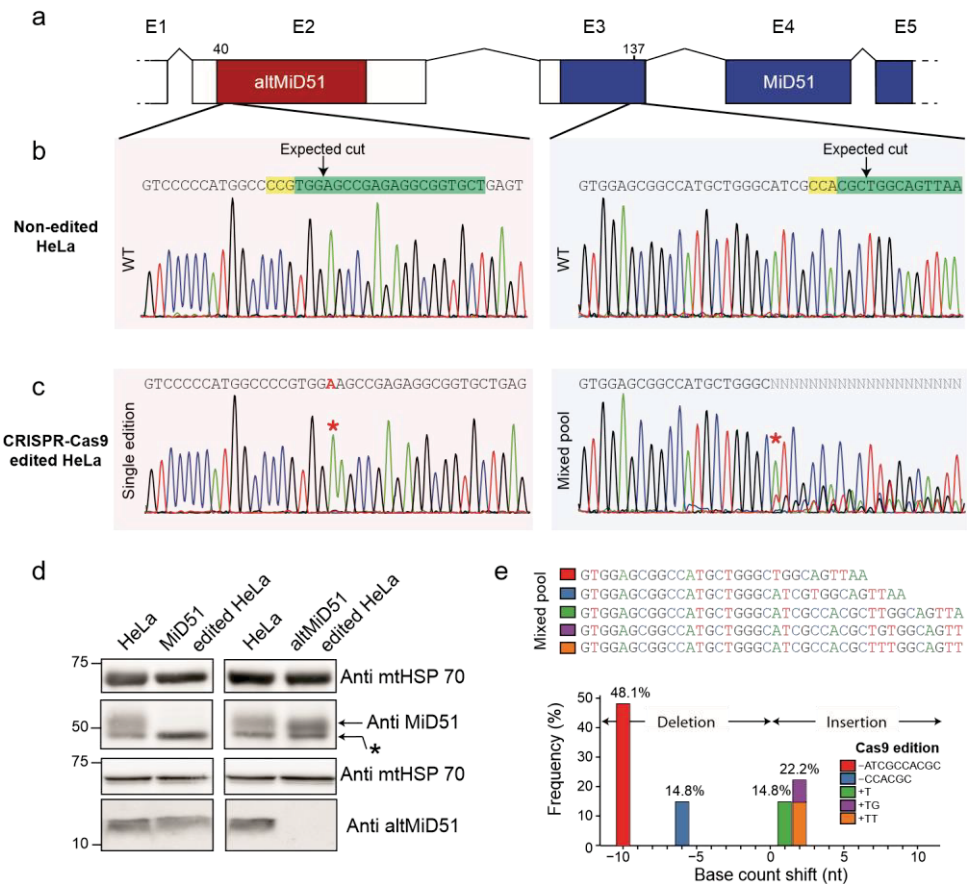

Figure 3

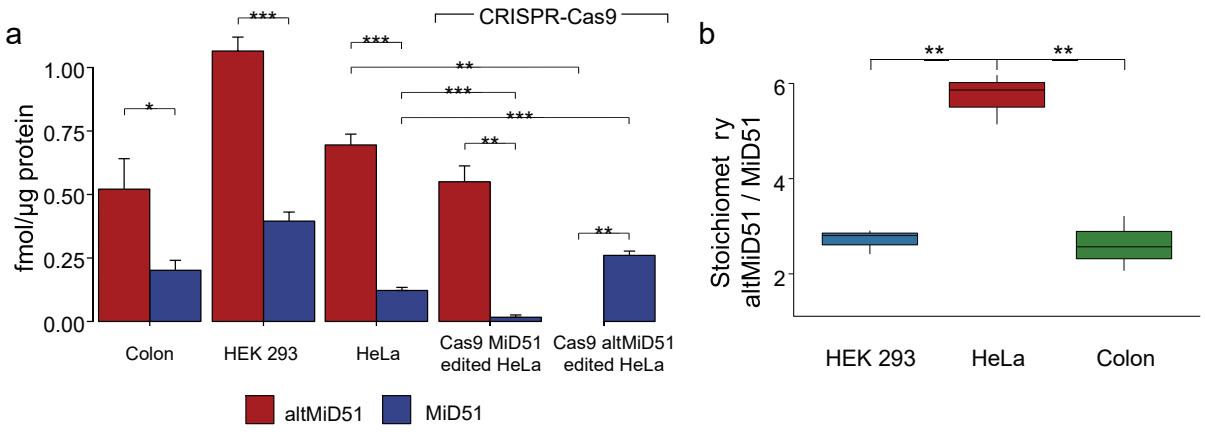

Figure 4