Deep transcriptome annotation suggests that small and large proteins encoded in the same genes often cooperate

Sondos Samandi[1,7]†, Annie V. Roy[1,7]†, Vivian Delcourt[1,7,8], Jean-François Lucier[2], Jules Gagnon[2], Maxime C. Beaudoin[1,7], Benoît Vanderperre[1], Marc-André Breton[1], Julie Motard[1,7], Jean-François Jacques[1,7], Mylène Brunelle[1,7], Isabelle Gagnon-Arsenault[6,7], Isabelle Fournier[8], Aida Ouangraoua[3], Darel J. Hunting[4], Alan A. Cohen[5], Christian R. Landry[6,7], Michelle S. Scott[1], Xavier Roucou[1,7]*

[1]Department of Biochemistry, [2]Department of Biology and Center for Computational Science, [3]Department of Computer Science, [4]Department of Nuclear Medicine & Radiobiology, [5]Department of Family Medicine, Université de Sherbrooke, Quebec, Canada; [6]Département de biologie and IBIS, Université Laval, Quebec, Canada; [7]PROTEO, Quebec Network for Research on Protein Function, Structure, and Engineering, Quebec, Canada; [8]Univ. Lille, INSERM U1192, Laboratoire Protéomique, Réponse Inflammatoire & Spectrométrie de Masse (PRISM) F-59000 Lille, France

†These authors contributed equally to this work

*Correspondance to Xavier Roucou: Department of Biochemistry (Z8-2001), Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, 3201 Jean Mignault, Sherbrooke, Quebec J1E 4K8, Canada, Tel. (819) 821-8000x72240; Fax. (819) 820 6831; E-Mail: xavier.roucou@usherbrooke.ca

Abstract

Recent studies in eukaryotes have demonstrated the translation of alternative open reading frames (altORFs) in addition to annotated protein coding sequences (CDSs). We show that a large number of small proteins could in fact be coded by altORFs. The putative alternative proteins translated from altORFs have orthologs in many species and evolutionary patterns indicate that altORFs are particularly constrained in CDSs that evolve slowly. Thousands of predicted alternative proteins are detected in proteomic datasets by reanalysis with a database containing predicted alternative proteins. Protein domains and co-conservation analyses suggest potential functional cooperation or shared function between small and large proteins encoded in the same genes. This is illustrated with specific examples, including altMID51, a 70 amino acid mitochondrial fission-promoting protein encoded in MiD51/Mief1/SMCR7L, a gene encoding an annotated protein promoting mitochondrial fission. Our results suggest that many coding genes code for more than one protein that are often functionally related.

Introduction

Current protein databases are cornerstones of modern biology but are based on a number of assumptions. In particular, a mature mRNA is predicted to contain a single CDS; yet, ribosomes can select more than one translation initiation site (TIS)[1–3] on any single mRNA. Also, minimum size limits are imposed on the length of CDSs, resulting in many RNAs being mistakenly classified as non-coding (ncRNAs)[4–11]. As a result of these assumptions, the size and complexity of most eukaryotic proteomes have probably been greatly underestimated[12–15]. In particular, few small proteins (defined as of 100 amino acids or less) are annotated in current databases. The absence of annotation of small proteins is a major bottleneck in the study of their function and to a full understanding of cell biology in health and disease. This is further supported by classical and recent examples of small proteins of functional importance, for instance many critical regulatory molecules such as F0 subunits of the F0F1-ATPsynthase[16], the sarcoplasmic reticulum calcium ATPase regulator phospholamban[17], and the key regulator of iron homeostasis hepcidin[18]. This limitation also impedes our understanding of the process of origin of genes *de novo*, which are thought to contribute to evolutionary innovations. Because these genes generally code for small proteins [19–22], they are difficult to detect by proteomics or even impossible to detect if they are not included in proteomics databases.

Functional annotation of ORFs encoding small proteins is particularly challenging since an unknown fraction of small ORFs may occur by chance in the transcriptome, generating a significant level of noise [13]. However, given that many small proteins have

63 important functions and are ultimately one of the most important sources of functional
64 novelty, it is time to address the challenge of their functional annotations[13].
65
66 We systematically reanalyzed several eukaryotic transcriptomes to annotate previously
67 unannotated ORFs which we term alternative ORFs (altORFs), and we annotated the
68 corresponding hidden proteome. Here, altORFs are defined as potential protein-coding
69 ORFs in ncRNAs or exterior to, or in different reading frames from annotated CDSs in
70 mRNAs (Figure 1a). For clarity, predicted proteins translated from altORFs are termed
71 alternative proteins and proteins translated from annotated CDSs are termed reference
72 proteins.
73
74 Our goal was to provide functional annotations of alternative proteins by (1) analyzing
75 relative patterns of evolutionary conservation between alternative and reference proteins
76 and their corresponding coding sequences;(2) estimating the prevalence of alternative
77 proteins both by bioinformatics analysis and by detection in large experimental datasets;
78 (3) detecting functional signatures in alternative proteins; and (4) predicting and testing
79 functional cooperation between alternative and reference proteins.
80

81 ## Results

82 **Prediction of altORFs and alternative proteins.** We predicted a total of 551,380
83 altORFs compared to 67,765 annotated CDSs in the human transcriptome (Figure 1b,
84 Table 1). Because identical ORFs can be present in different RNA isoforms transcribed
85 from the same genomic locus, the number of unique altORFs and CDSs becomes 183,191

and 51,818, respectively. AltORFs were also predicted in other organisms for comparison (Table 1). By convention, only reference proteins are annotated in current protein databases. As expected, these altORFs are on average small, with a size ranging from 30 to 1480 codons. Accordingly, the median size of human predicted alternative proteins is 45 amino acids compared to 460 for reference proteins (Figure 1c), and 92.96 % of alternative proteins have less than 100 amino acids. Thus, the bulk of the translation products of altORFs would be small proteins. The majority of altORFs either overlap annotated CDSs in a different reading frame (35.98%) or are located in 3'UTRs (40.09%) (Figure 1d). Only about 10% of altORFs are located in repeat sequences (Figure 1-figure supplement 1). To assess whether observed altORFs could be attributable solely to random occurrence, due for instance to the base composition of the transcriptome, we estimated the expected number of altORFs generated in 100 shuffled human transcriptomes. Overall, we observed 62,307 more altORFs than would be expected from random occurrence alone (Figure 1e; $p < 0.0001$). This analysis suggests that a large number are expected by chance alone but that at the same time, a large absolute number could potentially be maintained and be functional. The density of altORFs observed in the CDSs, 3'UTRs and ncRNAs (Figure 1f) was markedly higher than in the shuffled transcriptomes, suggesting that these are maintained at frequency higher than expected by chance, again potentially due to their coding function. In contrast, the density of altORFs observed in 5'UTRs was much lower than in the shuffled transcriptomes, supporting recent claims that negative selection eliminates AUGs (and thus the potential for the evolution of altORFs) in these regions[23,24].

Although the majority of human annotated CDSs do not have a TIS with a Kozak motif (Figure 1g)[25], there is a correlation between a Kozak motif and translation efficiency[26]. We find that 27,539 (15% of 183,191) human altORFs encoding predicted alternative proteins have a Kozak motif, as compared to 19,745 (38% of 51,818) for annotated CDSs encoding reference proteins (Figure 1g). The number of altORFs with Kozak motifs is significantly higher in the human transcriptome compared to shuffled transcriptomes (Figure 1-figure supplement 2), again supporting their potential role as protein coding.

Conservation analyses. Next, we compared evolutionary conservation patterns of altORFs and CDSs. A large number of human alternative proteins have homologs in other species. In mammals, the number of homologous alternative proteins is higher than the number of homologous reference proteins (Figure 2a), and 9 are even conserved from human to yeast (Figure 2b), supporting a potential functional role. As phylogenetic distance from human increases, the number and percentage of genes encoding homologous alternative proteins decreases more rapidly than the percentage of genes encoding reference proteins (Figure 2a, 2c). This observation indicates either that altORFs evolve more rapidly than CDSs or that distant homologies are less likely to be detected given the smaller sizes of alternative proteins. Another possibility is that they evolve following the patterns of evolution of genes that evolve *de novo*, with a rapid birth and death rate, which accelerates their turnover over time[20].

If altORFs play a functional role, they would be expected to be under purifying selection. The first and second positions of a codon experience stronger purifying selection than the

third[27]. By definition, CDS regions overlapping altORFs with a shifted reading frame do not contain such third positions because the third codon positions of the CDSs are either the first or the second in the altORFs. We analyzed conservation of third codon positions of CDSs for 100 vertebrate species for the 53,862 altORFs completely nested within the 20,814 CDSs from 14,677 genes (Figure 3). We observed that in regions of the CDS overlapping altORFs, third codon positions were evolving significantly more slowly than third codon positions of random control sequences from the entire CDS for a large number of altORFs (Figure 3), reaching up to 22-fold for conservation at $p<0.0001$. This is illustrated with three altORFs located within the CDS of NTNG1, RET and VTI1A genes (Figure 4). These three genes encode a protein promoting neurite outgrowth, the proto-oncogene tyrosine-protein kinase receptor Ret and a protein mediating vesicle transport to the cell surface, respectively. Two of these alternative proteins have been detected by ribosome profiling (RET, IP_182668.1) or mass spectrometry (VTI1A, IP_188229.1) (see below, supplementary files 1 and 2).

Evidence of expression of alternative proteins. We provide two lines of evidence indicating that thousands of altORFs are translated into proteins. First, we re-analyzed detected TISs in publicly available ribosome profiling data[28,29], and found 26,531 TISs mapping to annotated CDSs and 12,616 mapping to altORFs in these studies (Figure 5a; Supplementary file 1). Although predicted altORFs$^{3'}$ are more abundant than altORFs$^{5'}$, only a small fraction of TISs detected by ribosomal profiling mapped to altORFs$^{3'}$. Only a small fraction of TISs detected by ribosomal profiling mapped to altORFs$^{3'}$ even if those are more abundant than altORF$^{5'}$ relative to shuffled transcriptomes, likely reflecting a

recently-resolved technical issue in the ribosome profiling technique [30]. New methods to analyze ribosome profiling data are being developed and will likely uncover more translated altORFs [9]. In agreement with the presence of functional altORFs[3'], cap-independent translational sequences were recently discovered in human 3'UTRs[31]. New methods to analyze ribosome profiling data are being developed and will likely uncover more translated altORFs[9]. Second, we re-analyzed proteomic data using our composite database containing alternative proteins in addition to annotated reference proteins (Figure 5b, Supplementary file 2). False discovery rate cut-offs were set at 1% for peptide-spectrum match, peptides and proteins. We selected four studies representing different experimental paradigms and proteomic applications: large-scale [32] and targeted [33] protein/protein interactions, post-translational modifications [34], and a combination of bottom-up, shotgun and interactome proteomics [35] (Figure 5b). In the first dataset, we detected 7,530 predicted alternative proteins in the interactome of reference proteins[32], providing a framework to uncover the function of these proteins. In a second proteomic dataset containing about 10,000 reference human proteins[35], a total of 1,658 predicted alternative proteins were detected, representing more than 10% of the detectable proteome. Using a phosphoproteomic large data set[34], we detected 1,424 alternative proteins. The biological function of these proteins is supported by the observation that some alternative proteins are specifically phosphorylated in cells stimulated by the epidermal growth factor, and others are specifically phosphorylated during mitosis (Figure 6; Supplementary file 3). We provide examples of spectra validation using synthetic peptides (Figure 6-figure supplement 1-2). A fourth proteomic dataset contained 113 alternative proteins in the epidermal growth factor receptor interactome[33] (Figure

5b). A total of 10,362 different alternative proteins were detected in these proteomic data. Overall, by mining the proteomic and ribosomal profiling data, we detected the translation of a total of 22,155 unique alternative proteins. 823 of these alternative proteins were detected by both MS and ribosome profiling (Figure 7), providing a high-confidence collection of nearly one thousand small alternative proteins for further studies.

Functional annotations of alternative proteins. An important goal of this study is to associate potential functions to alternative proteins, which we can do through annotations. Because the sequence similarities and the presence of particular signatures (families, domains, motifs, sites) are a good indicator of a protein's function, we analyzed the sequence of the predicted alternative proteins in several organisms with InterProScan, an analysis and classification tool for characterizing unknown protein sequences by predicting the presence of combined protein signatures from most main domain databases[36] (Figure 8; Figure 8-figure supplement 1). We found 41,511 (23%) human alternative proteins with at least one InterPro signature (Figure 8b). Of these, 37,739 (or 20.6%) are classified as small proteins. Interestingly, the reference proteome has a smaller proportion (840 or 1.6%) of small proteins with at least one InterPro signature, supporting a biological activity for alternative proteins.

Similar to reference proteins, signatures linked to membrane proteins are abundant in the alternative proteome and represent more than 15,000 proteins (Figure 8c-e; Figure 8-supplemental figure 1). With respect to the targeting of proteins to the secretory pathway or to cellular membranes, the main difference between the alternative and the reference proteomes lies in the very low number of proteins with both signal peptides and

transmembrane domains. Most of the alternative proteins with a signal peptide do not have a transmembrane segment and are predicted to be secreted (Figure 8c, d), supporting the presence of large numbers of alternative proteins in plasma[37]. The majority of predicted alternative proteins with transmembrane domains have a single membrane spanning domain but some display up to 27 transmembrane regions, which is still within the range of reference proteins that show a maximum of 33 (Figure 8e).

A total of 585 alternative proteins were assigned 419 different InterPro entries, and 343 of them were tentatively assigned 192 gene ontology terms (Figure 9). 17.1% (100/585) of alternative proteins with an InterPro entry were detected by MS or/and ribosome profiling, compared to 13.7% (22,055/161,110) for alternative proteins without an InterPro entry. Thus, predicted alternative proteins with InterPro entries are more likely to be detected, supporting their functional role (p-value = 0.000035, Fisher's exact test and chi-square test). The most abundant class of predicted alternative proteins with at least one InterPro entry are C2H2 zinc finger proteins with 110 alternative proteins containing 187 C2H2-type/integrase DNA-binding domains, 91 C2H2 domains and 23 C2H2-like domains (Figure 10a). Seventeen of these (15.4%) were detected in public proteomic and ribosome profiling datasets, a percentage that is similar to reference zinc finger proteins (20.1%) (Figure 2, Table 2). Alternative proteins have between 1 and 23 zinc finger domains (Figure 10b). Zinc fingers mediate protein-DNA, protein-RNA and protein-protein interactions[38]. The linker sequence separating adjacent finger motifs matches or resembles the consensus TGEK sequence in nearly half the annotated zinc finger proteins[39]. This linker confers high affinity DNA binding and switches from a flexible to a rigid conformation to stabilize DNA binding. The consensus TGEK linker is present 46

times in 31 alternative zinc finger proteins (Supplementary file 4). These analyses show that a number of alternative proteins can be classified into families and will help deciphering their functions.

Evidence of functional coupling between reference and alternative proteins coded by the same genes. Since one gene codes for both a reference and one or several alternative proteins, we asked whether paired (encoded in the same gene) alternative and reference proteins have functional relationships. There are a few known examples of functional interactions between different proteins encoded in the same gene (Table 3). If there is functional cooperation or shared function, one would expect orthologous alternative-reference protein pairs to be co-conserved[40]. Our results show a large fraction of co-conserved alternative- reference protein pairs in several species (Figure 11). Detailed results for all species are presented in Table 4.

Another mechanism that could functionally associate alternative and reference proteins from the same transcripts would be that they share protein domains. We compared the functional annotations of the 585 alternative proteins with an InterPro entry with the reference proteins expressed from the same genes. Strikingly, 89 of 110 altORFs coding for zinc finger proteins (Figure 10) are present in transcripts in which the CDS also codes for a zinc finger protein. Overall, 138 alternative/reference protein pairs share at least one InterPro entry and many pairs share more than one entry (Figure 12a). The number of shared entries was much higher than expected by chance (Figure 12b, $p<0.0001$). The correspondence between InterPro domains of alternative proteins and their corresponding reference proteins coded by the same transcripts also indicates that even when entries are

not identical, the InterPro terms are functionally related (Figure 12c; Figure 12-figure supplement 1), overall supporting a potential functional association between reference and predicted alternative proteins. Domain sharing remains significant even when the most frequent domains, zinc fingers, are not considered (Figure 12-figure supplement 2).

Recently, the interactome of each of 131 human zinc finger proteins was determined by affinity purification followed by mass spectrometry [41]. This study provides a unique opportunity to test if, in addition to posessing zinc finger domains, some pairs of reference and alternative proteins coded by the same gene also interact. We re-analyzed the MS data using our alternative protein sequence database to detect alternative proteins in this interactome. Five alternative proteins were identified within the interactome of their reference zinc finger proteins. This number was higher than expected by chance ($p<10^{-6}$) based on 1 million binomial simulations of randomized interactomes. This result strongly supports the hypothesis of functional cooperation between alternative and reference proteins coded by the same genes.

Finally, we integrated the co-conservation and expression analyses to produce a high-confidence list of predicted functional and co-operating alternative proteins and found 3,028 alternative proteins in mammals (*H. sapiens* to *B. taurus*), and 51 in vertebrates (H. *sapiens* to *D. rerio*) (supplementary file 6). In order to further test for functional cooperation between alternative/reference protein pairs in this list, we focused on alternative proteins detected with at least two peptide spectrum matches. From this subset, we selected altMID51 (IP_294711.1) among the top 3% of alternative proteins

detected with the highest number of peptide spectrum matches in proteomics studies, and altDDIT3 (IP_211724.1) among the top 3% of altORFs with the most cumulative reads in translation initiation ribosome profiling studies.

AltMiD51 is a 70 amino acid alternative protein conserved in vertebrates[42] and co-conserved with its reference protein MiD51 from humans to zebrafish (supplementary file 6). Its coding sequence is present in exon 2 of the *MiD51/MIEF1/SMCR7L* gene. This exon forms part of the 5'UTR for the canonical mRNA and is annotated as non-coding in current gene databases (Figure 13a). Yet, altMID51 is robustly detected by MS in several cell lines (Supplementary file 2: HEK293, HeLa, HeLa S3, LNCaP, NCI60 and U2OS cells), and we validated some spectra using synthetic peptides (Figure 13-figure supplement 1), and is also detected by ribosome profiling (Supplementary file 1)[37,42,43]. We confirmed co-expression of altMiD51 and MiD51 from the same transcript (Figure 13b). Importantly, the tripeptide LYR motif predicted with InterProScan and located in the N-terminal domain of altMiD51 (Figure 13a) is a signature of mitochondrial proteins localized in the mitochondrial matrix[44]. Since *MiD51/MIEF1/SMCR7L* encodes the mitochondrial protein MiD51, which promotes mitochondrial fission by recruiting cytosolic Drp1, a member of the dynamin family of large GTPases, to mitochondria[45], we tested for a possible functional connection between these two proteins expressed from the same mRNA. We first confirmed that MiD51 induces mitochondrial fission (Figure 13-figure supplement 2). Remarkably, we found that altMiD51 also localizes at the mitochondria (Figure 13c; Figure 13-figure supplement 3) and that its overexpression results in mitochondrial fission (Figure 13d). This activity is unlikely to be through perturbation of oxidative phosphorylation since the overexpression of altMiD51 did not

293 change oxygen consumption nor ATP and reactive oxygen species production (Figure 13-
294 figure supplement 4). The decrease in spare respiratory capacity in altMiD51-expressing
295 cells (Figure 13-figure supplement 4a) likely resulted from mitochondrial fission[46]. The
296 LYR domain is essential for altMiD51-induced mitochondrial fission since a mutant of
297 the LYR domain, altMiD51(LYR→AAA) was unable to convert the mitochondrial
298 morphology from tubular to fragmented (Figure 13d). Drp1(K38A), a dominant negative
299 mutant of Drp1 [47], largely prevented the ability of altMiD51 to induce mitochondrial
300 fragmentation (Figure 13d; Figure 13-figure supplement 5a). In a control experiment, co-
301 expression of wild-type Drp1 and altMiD51 proteins resulted in mitochondrial
302 fragmentation (Figure 13-figure supplement 5b). Expression of the different constructs
303 used in these experiments was verified by western blot (Figure 13-figure supplement 6).
304 Drp1 knockdown interfered with altMiD51-induced mitochondrial fragmentation (Figure
305 14), confirming the proposition that Drp1 mediates altMiD51-induced mitochondrial
306 fragmentation. It remains possible that altMiD51 promotes mitochondrial fission
307 independently of Drp1 and is able to reverse the hyperfusion induced by Drp1
308 inactivation. However, Drp1 is the key player mediating mitochondrial fission and most
309 likely mediates altMiD51-induced mitochondrial fragmentation, as indicated by our
310 results.
311 AltDDIT3 is a 34 amino acid alternative protein conserved in vertebrates and co-
312 conserved with its reference protein DDIT3 from human to bovine (supplementary file
313 6). Its coding sequence overlaps the end of exon 1 and the beginning of exon 2 of the
314 *DDIT3/CHOP/GADD153* gene. These exons form part of the 5'UTR for the canonical
315 mRNA (Figure 15a). To determine the cellular localization of altDDIT3 and its possible

relationship with DDIT3, confocal microscopy analyses were performed on HeLa cells co-transfected with altDDIT3^{GFP} and DDIT3^{mCherry}. Interestingly, both proteins were mainly localized in the nucleus and partially localized in the cytoplasm (Figure 15b). This distribution for DDIT3 confirms previous studies [48,49]. Both proteins seemed to co-localize in these two compartments (Pearson correlation coefficient of 0.92, Figure 15c). We further confirmed the statistical significance of this colocalization by applying Costes' automatic threshold and Costes' randomization colocalization analysis and Manders Correlation Coefficient (Figure 15d) [50]. This was tested by co-immunoprecipitation. In lysates from cells co-expressing altDDIT3^{GFP} and DDIT3^{mCherry}, DDIT3^{mCherry} was immunoprecipitated with anti-GFP antibodies, confirming an interaction between the small altDDTI3 and the large DDIT3 proteins encoded in the same gene.

Discussion

In light of the increasing evidence from approaches such as ribosome profiling and MS-based proteomics that the one mRNA-one canonical CDS assumption is strongly challenged, our findings provide the first clear functional insight into a new layer of regulation in genome function. While many observed altORFs may be evolutionary accidents with no functional role, at least 9 independent lines of evidence support translation and a functional role for thousands of alternative proteins: (1) overrepresentation of altORFs relative to shuffled sequences; (2) overrepresentation of altORF Kozak sequences; (3) active altORF translation detected via ribosomal profiling;

(4) detection of thousand alternative proteins in multiple existing proteomic databases; (5) correlated altORF-CDS conservation, but with overrepresentation of highly conserved and fast-evolving altORFs; (6) underrepresentation of altORFs in repeat sequences; (7) overrepresentation of identical InterPro signatures between alternative and reference proteins encoded in the same mRNAs; (8) several thousand co-conserved paired alternative-reference proteins encoded in the same gene; and (9) presence of clear, striking examples in altMiD51, altDDI3T and 5 alternative proteins interacting with their reference zinc finger proteins. While 5 of these 9 lines of evidence support an unspecified functional altORF role, 4 of them (5, 7, 8 and 9) independently support a specific functional/evolutionary interpretation of their role: that alternative proteins and reference proteins have paired functions. Note that this hypothesis does not require binding, just functional cooperation such as activity on a shared pathway.

Upstream ORFs here labeled altORFs$^{5'}$ are important translational regulators of canonical CDSs in vertebrates[51]. Interestingly, the altORF5' encoding altDDIT3 was characterized as an inhibitory upstream ORF [52], but the corresponding small protein was not sought. The detection of altMiD51 and altDDI3T suggests that a fraction of altORFs$^{5'}$ may have dual functions as translation regulators and functional proteins.

Our results raise the question of the evolutionary origins of these altORFs. A first possible mechanism involves the polymorphism of initiation and stop codons during evolution [53,54]. For instance, the generation of an early stop codon in the 5'end of a CDS could be followed by the evolution of another translation initiation site downstream,

creating a new independent ORF in the 3'UTR of the canonical gene. This mechanism of altORF origin, reminiscent of gene fission, would at the same time produce a new altORF that shares protein domains with the annotated CDS, as we observed for a substantial fraction (24%) of the 585alternative proteins with an InterPro entry. A second mechanism would be de novo origin of ORFs, which would follow the well-established models of gene evolution *de novo*[20,55,56] in which new ORFs are transcribed and translated and have new functions or await the evolution of new functions by mutations. The numerous altORFs with no detectable protein domains may have originated this way from previously non-coding regions or in regions that completely overlap with CDS in other reading frames.

Detection is an important challenge in the study of small proteins. A TIS detected by ribosome profiling does not necessarily imply that the protein is expressed as a stable molecule, and proteomic analyses more readily detect large proteins that generate several peptides after enzymatic digestion. In addition, evolutionary novel genes tend to be lowly expressed, again reducing the probability of detection [20]. Here, we used a combination of five search engines and false discovery rate cut-offs were set at 1% for peptide-spectrum match, peptides and proteins, thus increasing the confidence and sensitivity of hits compared to single-search-engine processing[57,58]. This strategy led to the detection of several thousand alternative proteins. However, ribosome profiling and MS have technical caveats and the comprehensive contribution of small proteins to the proteome will require more efforts, including the development of new tools such as specific antibodies.

In conclusion, our deep annotation of the transcriptome reveals that a large number of small eukaryotic proteins, which may even represent the majority, are still officially unannotated. Our results also suggest that many small and large proteins coded by the same mRNA may cooperate by regulating each other's function or by functioning in the same pathway, confirming the few examples in the literature of unrelated proteins encoded in the same genes and functionally cooperating[59–63]. To determine whether or not this functional cooperation is a general feature of small/large protein pairs encoded in the same gene will require much more experimental evidence, but our results strongly support this hypothesis.

Materials and methods

Generation of alternative open reading frames (altORFs) and alternative protein databases. Throughout this manuscript, annotated protein coding sequences and proteins in current databases are labelled annotated coding sequences or CDSs and reference proteins, respectively. For simplicity reasons, predicted alternative protein coding sequences are labelled alternative open reading frames or altORFs.

To generate MySQL databases containing the sequences of all predicted alternative proteins translated from reference annotation of different organisms, a computational pipeline of Perl scripts was developed as previously described with some modifications[37]. Genome annotations for *H. sapiens* (release hg38, Assembly: GCF_000001405.26), *P. troglodytes* (Pan_troglodytes-2.1.4, Assembly: GCF_000001515.6), *M. musculus* (GRCm38.p2, Assembly: GCF_000001635.22), *D. melanogaster* (release 6, Assembly:

GCA_000705575.1), *C. elegans* (WBcel235, Assembly: GCF_000002985.6) and *S. cerevisiae* (Sc_YJM993_v1, Assembly: GCA_000662435.1) were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/genome). For *B. taurus* (release UMD 3.1.86), *X. tropicalis* (release JGI_4.2) and *D. rerio* (GRCz10.84), genome annotations were downloaded from Ensembl (http://www.ensembl.org/info/data/ftp/). Each annotated transcript was translated *in silico* with Transeq[64]. All ORFs starting with an AUG and ending with a stop codon different from the CDS, with a minimum length of 30 codons (including the stop codon) and identified in a distinct reading frame compared to the annotated CDS were defined as altORFs.

An additional quality control step was performed to remove initially predicted altORFs with a high level of identity with reference proteins. Such altORFs typically start in a different coding frame than the reference protein but through alternative splicing, end with the same amino acid sequence as their associated reference protein. Using BLAST, altORFs overlapping CDSs chromosomal coordinates and showing more than 80% identity and overlap with an annotated CDS were rejected.

AltORF localization was assigned according to the position of the predicted translation initiation site (TIS): altORFs$^{5'}$, altORFs^{CDS} and altORFs$^{3'}$ are altORFs with TISs located in 5'UTRs, CDSs and 3'UTRs, respectively. Non-coding RNAs (ncRNAs) have no annotated CDS and all ORFs located within ncRNAs are labelled altORFs^{nc}.

The presence of the simplified Kozak sequence (A/GNNATGG) known to be favorable for efficient translation initiation was also assessed for each predicted altORF[65].

Identification of TISs. The global aggregates of initiating ribosome profiles data were

obtained from the initiating ribosomes tracks in the GWIPS-viz genome browser[28] with ribosome profiling data collected from five large scale studies[2,9,66–68]. Sites were mapped to hg38 using a chain file from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz) and CrossMap v0.1.6 (http://crossmap.sourceforge.net/). Similar to the methods used in these studies, an altORF is considered as having an active TIS if it is associated with at least 10 reads at one of the 7 nucleotide positions of the sequence NNNAUGN (AUG is the predicted altORF TIS). An additional recent study was also included in our analysis[29]. Raw sequencing data for ribosome protected fragments in harringtonine treated cells was aligned to the human genome (GRCh38) using bowtie2 (2.2.8). Similar to the method used in this work, altORFs with at least 5 reads overlapping one position in the kozak region were considered as having an experimentally validated TIS.

Generation of shuffled transcriptomes. Each annotated transcript was shuffled using the Fisher-Yates shuffle algorithm. In CDS regions, all codons were shuffled except the initiation and stop codons. For mRNAs, we shuffled the 5'UTRs, CDSs and 3'UTRs independently to control for base composition. Non-coding regions were shuffled at the nucleotide level. The resulting shuffled transcriptome has the following features compared to hg38: same number of transcripts, same transcripts lengths, same nucleotide composition, and same amino-acid composition for the proteins translated from the CDSs. Shuffling was repeated 100 times and the results are presented with average values and standard deviations. The total number of altORFs is 551,380 for hg38, and an average of 489,073 for shuffled hg38. AltORFs and kozak motifs in the 100 shuffled

transcriptomes were detected as described above for hg38.

Identification of paralogs/orthologs in alternative proteomes. Both alternative and reference proteomes were investigated. Pairwise ortholog and paralog relationships between the human proteomes and the proteomes from other species, were calculated using an InParanoid-like approach[69], as described below. The following BLAST procedure was used. Comparisons using our datasets of altORFs/CDS protein sequences in multiple FASTA formats from *Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, Xenopus tropicalis Bos taurus, Mus musculus, Pan troglodytes, Homo sapiens* were performed between each pair of species (human against the other species), involving four whole proteome runs per species pair: pairwise comparisons (organism A vs organism B, organism B vs organism A), plus two self-self runs(organism A vs organism A, organism B vs organism B). BLAST homology inference was accepted when the length of the aligned region between the query and the match sequence equalled or exceeded 50% of the length of the sequence, and when the bitscore reached a minimum of 40[70]. Orthologs were detected by finding the mutually best scoring pairwise hits (reciprocal best hits) between datasets A-B and B-A. The self-self runs were used to identify paralogy relationships as described[69].

Co-conservation analyses. For each orthologous alternative protein pair A-B between two species, we evaluated the presence and the orthology of their corresponding reference proteins A'-B' in the same species. In addition, the corresponding altORFs and CDSs had to be present in the same gene.

In order to develop a null model to assess co-conservation of alternative proteins and their reference pairs, we needed to establish a probability that any given orthologous alternative protein would by chance occur encoded on the same transcript as its paired, orthologous reference protein. Although altORFs might in theory shift among CDSs (and indeed, a few examples have been observed), transposition events are expected to be relatively rare; we thus used the probability that the orthologous alternative protein is paired with any orthologous CDS for our null model. Because this probability is by definition higher than the probability that the altORF occurs on the paired CDS, it is a conservative estimate of co-conservation. We took two approaches to estimating this percentage, and then used whichever was higher for each species pair, yielding an even more conservative estimate. First, we assessed the percentage of orthologous reference proteins under the null supposition that each orthologous alternative protein had an equal probability of being paired with any reference protein, orthologous or not. Second, we assessed the percentage of non-orthologous alternative proteins that were paired with orthologous reference proteins. This would account for factors such as longer CDSs having a higher probability of being orthologous and having a larger number of paired altORFs. For example, between humans and mice, we found that 22,304 of 51,819 reference proteins (43%) were orthologs. Of the 157,261 non-orthologous alternative proteins, 106,987 (68%) were paired with an orthologous reference protein. Because 68% is greater than 43%, we used 68% as the probability for use in our null model. Subsequently, our model strongly indicates co-conservation (Fig. 11; $p<10^{-6}$ based on 1 million binomial simulations; highest observed random percentage =69%, much lower than the observed 96% co-conservation).

Analysis of third codon position (wobble) conservation. Basewise conservation scores for the alignment of 100 vertebrate genomes including *H. sapiens* were obtained from UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP100way/). Conservation PhyloP scores relative to each nucleotide position within codons were extracted using a custom Perl script and the Bio-BigFile module version 1.07. The PhyloP conservation score for the wobble nucleotide of each codon within the CDS was extracted. For the 53,862 altORFs completely nested inside 20,814 CDSs, the average PhyloP score for wobble nucleotides within the altORF region was compared to the average score for the complete CDS. To generate controls, random regions in CDSs with a similar length distribution as altORFs were selected and PhyloP scores for wobble nucleotides were extracted. We compared the differences between altORF and CDS PhyloP scores (altORF PhyloP – CDS PhyloP) to those generated based on random regions. We identified expected quantiles of the differences ("DQ" column in the table), and compared these to the observed differences. Because there was greater conservation of wobble nucleotide PhyloP scores within altORFs regions located farther from the center of their respective genes ($r = 0.08$, $p < 0.0001$), observed differences were adjusted using an 8-knot cubic basis spline of percent distance from center. These observed differences were also adjusted for site-specific signals as detected in the controls.

Human alternative protein classification and in silico functional annotation.

Repeat and transposable element annotation

RepeatMasker, a popular software to scan DNA sequences for identifying and classifying repetitive elements, was used to investigate the extent of altORFs derived from transposable elements[71]. Version 3-3-0 was run with default settings.

Alternative protein analysis using InterProScan

InterProScan combines 15 different databases, most of which use Hidden Markov models for signature identification[72]. Interpro merges the redundant predictions into a single entry and provides a common annotation. A recent local version of InterProScan 5.14-53.0 was run using default parameters to scan for known protein domains in alternative proteins. Gene ontology (GO) and pathway annotations were also reported if available with -goterm and -pa options. Only protein signatures with an E-value $\leq 10^{-3}$ were considered.

We classified the reported InterPro hits as belonging to one or several of three clusters; (1) alternative proteins with InterPro entries; (2) alternative proteins with signal peptides (SP) and/or transmembrane domains (TM) predicted by at least two of the three SignalP, PHOBIUS, TMHMM tools and (3) alternative proteins with other signatures.

The GO terms assigned to alternative proteins with InterPro entries were grouped and categorised into 13 classes within the three ontologies (cellular component, biological process, molecular function) using the CateGOrizer tool[73].

Each unique alternative protein with InterPro entries and its corresponding reference protein (encoded in the same transcript) were retrieved from our InterProscan output. Alternative and reference proteins without any InterPro entries were ignored. The overlap in InterPro entries between alternative and reference proteins was estimated as follows. We went through the list of alternative/reference protein pairs and counted the overlap in

546 the number of entries between the alternative and reference proteins as
547 100*intersection/union. All reference proteins and the corresponding alternative proteins
548 were combined together in each comparison so that all domains of all isoforms for a
549 given reference protein were considered in each comparison. The random distribution of
550 the number of alternative/reference protein pairs that share at least one InterPro entry was
551 computed by shuffling the alternative/reference protein pairs and calculating how many
552 share at least one InterPro entry. This procedure was repeated 1,000 times. Finally, we
553 compared the number and identity of shared InterPro entries in a two dimensional matrix
554 to illustrate which Interpro entries are shared. In many instances, including for zinc-finger
555 coding genes, InterPro entries in alternative/reference protein pairs tend to be related
556 when they are not identical.
557
558 **Mass Spectrometry identification parameters.** Wrapper Perl scripts were developed for
559 the use of SearchGUI v2.0.11[74] and PeptideShaker v1.1.0[57] on the Université de
560 Sherbrooke's 39,168 core high-performance Mammouth Parallèle 2 computing cluster
561 (http://www.calculquebec.ca/en/resources/compute-servers/mammouth-parallele-ii).
562 SearchGUI was configured to run the following proteomics identification search engines:
563 X!Tandem[75], MS-GF+[76], MyriMatch[77], Comet[78], and OMSSA[79]. SearchGUI parameters
564 were set as follow: maximum precursor charge, 5; maximum number of PTM per peptide,
565 5; X!Tandem minimal fragment m/z, 140; removal of initiator methionine for Comet, 1. A
566 full list of parameters used for SearchGUI and PeptideShaker is available in
567 Supplementary file 2, sheet 1. For PXD000953 dataset[35], precursor and fragment
568 tolerance were set 0.006 Da and 0.1 Da respectively, with carbamidomethylation of C as

569 a fixed modification and Nter-Acetylation and methionine oxidation as variable
570 modifications. For PXD000788[33] and PXD000612[34] datasets, precursor and fragment
571 tolerance were set to 4.5 ppm and 0.1 Da respectively with carbamidomethylation of
572 cysteine as a fixed modification and Nter-Acetylation, methionine oxidation and
573 phosphorylation of serine, threonine and tyrosine as variable modifications. For
574 PXD002815 dataset[32], precursor and fragment tolerance were set to 4.5 ppm and 0.1 Da
575 respectively with carbamidomethylation of cysteine as a fixed modification and Nter-
576 Acetylation and methionine oxidation as variable modifications.Datasets were searched
577 using a target-decoy approach against a composite database composed of a target
578 database [Uniprot canonical and isoform reference proteome (16 January 2015) for a total
579 of 89,861 sequences + custom alternative proteome resulting from the in silico translation
580 of all human altORFs (available to download at
581 https://www.roucoulab.com/p/downloads)], and their reverse protein sequences from the
582 target database used as decoys. False discovery rate cut-offs were set at 1% for PSM,
583 peptides and proteins. Only alternative proteins identified with at least one unique and
584 specific peptide, and with at least one confident PSM in the PeptideShaker Hierarchical
585 Report were considered valid[57].
586 Peptides matching proteins in a protein sequence database for common contaminants
587 were rejected[80].
588 For spectral validation (Figure 13-figure supplement 1; Supplementary Figures 1-4),
589 synthetic peptides were purchased from the peptide synthesis service at the Université de
590 Sherbrooke. Peptides were solubilized in 10% acetonitrile, 1% formic acid and directly
591 injected into a Q-Exactive mass spectrometer (Thermo Scientific) via an electro spray

ionization source (Thermo Scientific). Spectra were acquired using Xcalibur 2.2 at 70000 resolution with an AGC target of 3e6 and HCD collision energy of 25. Peaks were assigned manually by comparing monoisotopic m/z theoretical fragments and experimental (PeptideShaker) spectra.

In order to test if the interaction between alternative zinc-finger/reference zinc-finger protein pairs (encoded in the same gene) may have occurred by chance only, all interactions between alternative proteins and reference proteins were randomized with an in-house randomisation script. The number of interactions with reference proteins for each altProt was kept identical as the number of observed interactions. The results indicate that interactions between alternative zinc-finger/reference zinc-finger protein pairs did not occur by chance ($p<10^{-6}$) based on 1 million binomial simulations; highest observed random interactions between alternative zinc-finger proteins and their reference proteins = 3 (39 times out of 1 million simulations), compared to detected interactions=5.

Code availability. Computer codes are available upon request with no restrictions.

Data availability. Most Data are available in Supplementary information. Alternative protein databases for different species can be accessed at https://www.roucoulab.com/p/downloads with no restrictions.

Cloning and antibodies. Human Flag-tagged altMiD51(WT) and altMiD51(LYR→AAA), and HA-tagged DrP1(K38A) were cloned into pcDNA3.1 (Invitrogen) using a Gibson assembly kit (New England Biolabs, E26115). The cDNA

614 corresponding to human MiD51/MIEF1/SMCR7L transcript variant 1 (NM_019008) was
615 also cloned into pcDNA3.1 by Gibson assembly. In this construct, altMiD51 and MiD51
616 were tagged with Flag and HA tags, respectively. MiD51^{GFP} and altMiD51^{GFP} were also
617 cloned into pcDNA3.1 by Gibson assembly. For MiD51^{GFP}, a LAP tag[32] was inserted
618 between MiD51 and GFP. gBlocks were purchased from IDT. Human altDDIT3^{mCherry}
619 was cloned into pcDNA3.1 by Gibson assembly using coding sequence from transcript
620 variant 1 (NM_001195053) and mCherry coding sequence from pLenti-myc-GLUT4-
621 mCherry (Addgene plasmid # 64049). Human DDIT3^{GFP} was also cloned into pcDNA3.1
622 by Gibson assembly using CCDS8943 sequence. gBlocks were purchased from IDT.
623 For immunofluorescence, primary antibodies were diluted as follow: anti-Flag (Sigma,
624 F1804) 1/1000, anti-TOM20 (Abcam, ab186734) 1/500. For western blots, primary
625 antibodies were diluted as follow: anti-Flag (Sigma, F1804) 1/1000, anti-HA (BioLegend,
626 901515) 1/500, anti-actin (Sigma, A5441) 1/10000, anti-Drp1 (BD Transduction
627 Laboratories, 611112) 1/500, anti-GFP (Santa Cruz Biotechnology, sc-9996) 1/10000,
628 anti-mCherry (Abcam, ab125096) 1/2000.
629
630 **Cell culture, immunofluorescence, knockdown and western blots.** HeLa cell (ATCC
631 CCL-2) cultures, transfections, immunofluorescence, confocal analyses and western blots
632 were carried out as previously described[81]. Mitochondrial morphology was analyzed as
633 previously described[82]. A minimum of 100 cells were counted (n=3 or 300 cells for each
634 experimental condition). Three independent experiments were performed.
635 For Drp1 knockdown, 25,000 HeLa cells in 24-well plates were transfected with 25 nM
636 Drp1 SMARTpool: siGENOME siRNA (Dharmacon, M-012092-01-0005) or ON-

637 TARGET plus Non-targeting pool siRNAs (Dharmacon, D-001810-10-05) with
638 DharmaFECT 1 transfection reagent (Dharmacon, T-2001-02) according to the
639 manufacturer's protocol. After 24h, cells were transfected with pcDNA3.1 or altMiD51,
640 incubated for 24h, and processed for immunofluorescence or western blot. Colocalization
641 analyses were performed using the JACoP plugin (Just Another Co-localization Plugin) [50]
642 implemented in Image J software.
643
644 **Mitochondrial localization, parameters and ROS production.** Trypan blue quenching
645 experiment was performed as previously described[83].
646 A flux analyzer (XF96 Extracellular Flux Analyzer; Seahorse Bioscience, Agilent
647 technologies) was used to determine the mitochondrial function in HeLa cells
648 overexpressing AltMiD51^{Flag}. Cells were plated in a XF96 plate (Seahorse Biosciences)
649 at 1×10^4 cells per well in Dulbecco's modified Eagle's medium supplemented with 10%
650 FBS with antibiotics. After 24 hours, cells were transfected for 24 hours with an empty
651 vector (pcDNA3.1) or with the same vector expressing AltMiD51^{Flag} with GeneCellin
652 tranfection reagent according to the manufacturer's instructions. Cells were equilibrated
653 in XF assay media supplemented with 25 mM glucose and 1 mM pyruvate and were
654 incubated at 37°C in a CO2-free incubator for 1h. Baseline oxygen consumption rates
655 (OCRs) of the cells were recorded with a mix/wait/measure times of 3/0/3 min
656 respectively. Following these measurements, oligomycin (1 µM), FCCP (0.5 µM), and
657 antimycin A/rotenone (1 µM) were sequentially injected, with oxygen consumption rate
658 measurements recorded after each injection. Data were normalized to total protein in each
659 well. For normalization, cells were lysed in the 96-well XF plates using 15 µl/well of

660 RIPA lysis buffer (1% Triton X-100, 1% NaDeoxycholate, 0.1% SDS, 1mM EDTA, 50
661 mM Tris-HCl pH7.5). Protein concentration was measured using the BCA protein assay
662 reagent (Pierce, Waltham, MA, USA).
663 Reactive oxygen species (ROS) levels were measured using Cellular ROS/Superoxide
664 Detection Assay Kit (Abcam #139476). HeLa cells were seeded onto 96-well black/clear
665 bottom plates at a density of 6,000 cells per well with 4 replicates for each condition.
666 After 24 hours, cells were transfected for 24 hours with an empty vector (pcDNA3.1) or
667 with the same vector expressing AltMiD51^{Flag} with GeneCellin according to the
668 manufacturer's instruction. Cells were untreated or incubated with the ROS inhibitor (N-
669 acetyl-L-cysteine) at 10mM for 1 hour. Following this, the cells were washed twice with
670 the wash solution and then labeled for 1 hour with the Oxidative Stress Detection
671 Reagent (green) diluted 1:1000 in the wash solution with or without the positive control
672 ROS Inducer Pyocyanin at 100µM. Fluorescence was monitored in real time. ROS
673 accumulation rate was measured between 1 to 3 hours following induction. After the
674 assay, total cellular protein content was measured using BCA protein assay reagent
675 (Pierce, Waltham, MA, USA) after lysis with RIPA buffer. Data were normalised for
676 initial fluorescence and protein concentration.
677 ATP synthesis was measured as previously described[84] in cells transfected for 24 hours
678 with an empty vector (pcDNA3.1) or with the same vector expressing AltMiD51^{Flag}.
679

680 **Acknowledgements**

681 This research was supported by CIHR grants MOP-137056 and MOP-136962 to X.R;
682 MOP-299432 and MOP-324265 to C.L; a Université de Sherbrooke institutional research

683 grant made possible through a generous donation by Merck Sharp & Dohme to X.R; a
684 FRQNT team grant 2015-PR-181807 to C.L. and X.R; Canada Research Chairs in
685 Functional Proteomics and Discovery of New Proteins to X.R, in Evolutionary Cell and
686 Systems Biology to C.L and in Computational and Biological Complexity to A.O; A.A.C
687 is supported by a CIHR New Investigator Salary Award; M.S.S is a recipient of a Fonds
688 de Recherche du Québec – Santé Research Scholar Junior 1 Career Award; V.D is
689 supported in part by fellowships from Région Nord-Pas de Calais and PROTEO; A.A.C,
690 D.J.H, M.S.S and X.R are members of the Fonds de Recherche du Québec Santé-
691 supported Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke. We
692 thank the staff from the Centre for Computational Science at the Université de
693 Sherbrooke, Compute Canada and Compute Québec for access to the Mammouth
694 supercomputer.

695

696

References

1. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).

2. Lee, S. S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2424-2432 (2012).

3. Mouilleron, H., Delcourt, V. & Roucou, X. Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res.* **44,** 14–23 (2015).

4. Pauli, A. *et al.* Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* **343,** 1248636–1248636 (2014).

5. Anderson, D. M. *et al.* A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160,** 595–606 (2015).

6. Zanet, J. *et al.* Pri sORF peptides induce selective proteasome-mediated protein processing. *Science* **349,** 1356–1358 (2015).

7. Nelson, B. R. *et al.* A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science (80-.).* **351,** 271–275 (2016).

8. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33,** 981–993 (2014).

9. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4,** e08890 (2015).

10. Prabakaran, S. *et al.* Quantitative profiling of peptides from RNAs classified as

noncoding. *Nat. Commun.* **5,** 5429 (2014).

11. Slavoff, S. a *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).

12. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **15,** 193–204 (2014).

13. Landry, C. R., Zhong, X., Nielly-Thibault, L. & Roucou, X. Found in translation: Functions and evolution of a recently discovered alternative proteome. *Curr. Opin. Struct. Biol.* **32,** 74–80 (2015).

14. Fields, A. P. *et al.* A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* **60,** 816–827 (2015).

15. Saghatelian, A. & Couso, J. P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.* **11,** 909–16 (2015).

16. Stock, D., Leslie, A. G. & Walker, J. E. Molecular architecture of the rotary motor in ATP synthase. *Science* **286,** 1700–1705 (1999).

17. Schmitt, J. P. *et al.* Dilated cardiomyopathy and heart failure caused by a mutation in phospholamban. *Science* **299,** 1410–1413 (2003).

18. Nemeth, E. *et al.* Hepcidin regulates cellular iron efflux by binding to ferroportin and inducing its internalization. *Science* **306,** 2090–2093 (2004).

19. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* 3–7 (2012). doi:10.1038/nature11184

20. Schlötterer, C. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet.* **31,** 215–9 (2015).

743 21. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what,
744 how and why. *Nat. Rev. Genet.* **17,** 567–578 (2016).
745 22. Sabath, N., Wagner, A. & Karlin, D. Evolution of viral proteins originated de novo
746 by overprinting. *Mol. Biol. Evol.* **29,** 3767–80 (2012).
747 23. Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent
748 5'untranslated mRNAs. *Gene* **349,** 97–105 (2005).
749 24. Neafsey, D. E. & Galagan, J. E. Dual modes of natural selection on upstream open
750 reading frames. *Mol. Biol. Evol.* **24,** 1744–51 (2007).
751 25. Smith, E. *et al.* Leaky ribosomal scanning in mammalian genomes: significance of
752 histone H4 alternative translation in vivo. *Nucleic Acids Res.* **33,** 1298–1308
753 (2005).
754 26. Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the
755 efficiency of translation and elongation. *Mol. Syst. Biol.* **10,** 770 (2014).
756 27. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of
757 nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–
758 121 (2010).
759 28. Michel, A. M. *et al.* GWIPS-viz: development of a ribo-seq genome browser.
760 *Nucleic Acids Res.* **42,** D859-864 (2014).
761 29. Raj, A. *et al.* Thousands of novel translated open reading frames in humans
762 inferred by ribosome footprint profiling. *Elife* **5,** 1–24 (2016).
763 30. Miettinen, T. P. & Björklund, M. Modified ribosome profiling reveals high
764 abundance of ribosome protected mRNA fragments derived from 3' untranslated
765 regions. *Nucleic Acids Res.* **43,** 1019–1034 (2015).

766 31. Weingarten-Gabbay, S. *et al.* Systematic discovery of cap-independent translation
767 sequences in human and viral genomes. *Science (80-.).* **351,** 1–24 (2016).
768 32. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions
769 Organized by Stoichiometries and Abundances. *Cell* **163,** 712–723 (2015).
770 33. Tong, J., Taylor, P. & Moran, M. F. Proteomic analysis of the epidermal growth
771 factor receptor (EGFR) interactome and post-translational modifications associated
772 with receptor endocytosis in response to EGF and stress. *Mol. Cell. Proteomics* **13,**
773 1644–1658 (2014).
774 34. Sharma, K. *et al.* Ultradeep Human Phosphoproteome Reveals a Distinct
775 Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Rep.* **8,** 1583–1594
776 (2014).
777 35. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by
778 SWATH-MS. *Sci. data* **1,** 140031 (2014).
779 36. Mitchell, A. *et al.* The InterPro protein families database: the classification
780 resource after 15 years. *Nucleic Acids Res.* **43,** D213-221 (2014).
781 37. Vanderperre, B. *et al.* Direct detection of alternative open reading frames
782 translation products in human significantly expands the proteome. *PLoS One* **8,**
783 e70698 (2013).
784 38. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc
785 finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 183–212 (2000).
786 39. Laity, J. H., Lee, B. M. & Wright, P. E. Zinc finger proteins: new insights into
787 structural and functional diversity. *Curr. Opin. Struct. Biol.* **11,** 39–46 (2001).
788 40. Karimpour-Fard, A., Detweiler, C. S., Erickson, K. D., Hunter, L. & Gill, R. T.

Cross-species cluster co-conservation: a new method for generating protein interaction networks. *Genome Biol.* **8,** R185 (2007).

41. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26,** 1742–1752 (2016).

42. Andreev, D. E. *et al.* Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4,** e03971 (2015).

43. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).

44. Angerer, H. Eukaryotic LYR Proteins Interact with Mitochondrial Protein Complexes. *Biology (Basel).* **4,** 133–150 (2015).

45. Losón, O. C., Song, Z., Chen, H. & Chan, D. C. Fis1, Mff, MiD49, and MiD51 mediate Drp1 recruitment in mitochondrial fission. *Mol. Biol. Cell* **24,** 659–667 (2013).

46. Motori, E. *et al.* Inflammation-Induced Alteration of Astrocyte Mitochondrial Dynamics Requires Autophagy for Mitochondrial Network Maintenance. *Cell Metab.* **18,** 844–859 (2013).

47. Smirnova, E., Shurland, D. L., Ryazantsev, S. N. & van der Bliek, A. M. A human dynamin-related protein controls the distribution of mitochondria. *J. Cell Biol.* **143,** 351–358 (1998).

48. Cui, K., Coutts, M., Stahl, J. & Sytkowski, A. J. Novel interaction between the transcription factor CHOP (GADD153) and the ribosomal protein FTE/S3a modulates erythropoiesis. *J. Biol. Chem.* **275,** 7591–6 (2000).

49. Chiribau, C.-B., Gaccioli, F., Huang, C. C., Yuan, C. L. & Hatzoglou, M.

812 Molecular symbiosis of CHOP and C/EBP beta isoform LIP contributes to
813 endoplasmic reticulum stress-induced apoptosis. *Mol. Cell. Biol.* **30,** 3722–31
814 (2010).
815 50. Bolte, S. & Cordelières, F. P. A guided tour into subcellular colocalization analysis
816 in light microscopy. *J. Microsc.* **224,** 213–32 (2006).
817 51. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent
818 translational repressors in vertebrates. *EMBO J.* (2016).
819 doi:10.15252/embj.201592759
820 52. Jousse, C. *et al.* Inhibition of CHOP translation by a peptide encoded by an open
821 reading frame localized in the chop 5'UTR. *Nucleic Acids Res.* **29,** 4341–51
822 (2001).
823 53. Lee, Y. C. G. & Reinhardt, J. A. Widespread Polymorphism in the Positions of
824 Stop Codons in Drosophila melanogaster. *Genome Biol. Evol.* **4,** 533–549 (2012).
825 54. Andreatta, M. E. *et al.* The Recent De Novo Origin of Protein C-Termini. *Genome*
826 *Biol. Evol.* **7,** 1686–701 (2015).
827 55. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding
828 genes. *Genome Res.* **19,** 1752–9 (2009).
829 56. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a
830 model of frequent de novo evolution. *BMC Genomics* **14,** 117 (2013).
831 57. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data
832 sets. *Nat. Biotechnol.* **33,** 22–24 (2015).
833 58. Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun
834 proteomic data improves peptide and protein identification rates and error

estimates. *Mol. Cell. Proteomics* **10,** M111.007690 (2011).

59. Quelle, D. E., Zindy, F., Ashmun, R. A. & Sherr, C. J. Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83,** 993–1000 (1995).

60. Abramowitz, J., Grenet, D., Birnbaumer, M., Torres, H. N. & Birnbaumer, L. XLalphas, the extra-long form of the alpha-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 8366–8371 (2004).

61. Bergeron, D. *et al.* An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J. Biol. Chem.* **288,** 21824–35 (2013).

62. Lee, C. -f. C., Lai, H.-L. H.-L., Lee, Y.-C., Chien, C.-L. C.-L. & Chern, Y. The A2A Adenosine Receptor Is a Dual Coding Gene: A NOVEL MECHANISM OF GENE USAGE AND SIGNAL TRANSDUCTION. *J. Biol. Chem.* **289,** 1257–1270 (2014).

63. Yosten, G. L. C. *et al.* A 5′-Upstream short open reading frame encoded peptide regulates angiotensin type 1a receptor production and signaling via the beta-arrestin pathway. *J. Physiol.* **6,** n/a-n/a (2015).

64. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).

65. Kozak, M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299,** 1–34 (2002).

66. Fritsch, C. *et al.* Genome-wide search for novel human uORFs and N-terminal

protein extensions using ribosomal footprinting. *Genome Res.* **22,** 2208–2218 (2012).

67. Stern-Ginossar, N. *et al.* Decoding human cytomegalovirus. *Science* **338,** 1088–93 (2012).

68. Gao, X. *et al.* Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* **12,** 147–53 (2015).

69. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43,** D234-239 (2015).

70. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314,** 1041–1052 (2001).

71. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4,** Unit 4.10 (2009).

72. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).

73. Na, D., Son, H. & Gsponer, J. Categorizer: a tool to categorize genes into user-defined biological groups based on semantic similarity. *BMC Genomics* **15,** 1091 (2014).

74. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11,** 996–999 (2011).

75. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra.

Bioinformatics **20,** 1466–1467 (2004).

76. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5,** 5277 (2014).

77. Tabb, D. L., Fernando, C. G. & Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6,** 654–661 (2007).

78. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13,** 22–24 (2013).

79. Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3,** 958–64 (2004).

80. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).

81. Vanderperre, B. *et al.* An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB J.* **25,** 2373–86 (2011).

82. Palmer, C. S. *et al.* MiD49 and MiD51, new components of the mitochondrial fission machinery. *EMBO Rep.* **12,** 565–573 (2011).

83. Vanderperre, B. *et al.* MPC1-like: a Placental Mammal-Specific Mitochondrial Pyruvate Carrier Subunit Expressed in Post-Meiotic Male Germ Cells. *J. Biol. Chem.* (2016). doi:10.1074/jbc.M116.733840

84. Vives-Bauza, C., Yang, L. & Manfredi, G. Assay of Mitochondrial ATP Synthesis in Animal Cells and Tissues. *Methods Cell Biol* **80,** 155–171 (2007).

Supplementary figure 1: Spectra validation for altSLC35A4$^{5'}$

Supplementary Figure 2: Spectra validation for altRELT$^{5'}$

Supplementary Figure 3: Spectra validation for altLINC01420^{nc}

Supplementary Figure 4: Spectra validation for altSRRM2^{CDS}

Supplementary file 1: 12,616 alternative proteins with translation initiation sites detected by ribosome profiling after re-analysis of large scale studies. Sheet 1: list of alternative proteins; sheet 2: pie chart of corresponding altORFs localization.

Supplementary file 2: 10,362 alternative proteins detected by mass spectrometry (MS) after re-analysis of large proteomic studies. Sheet 1: MS identification parameters; sheet 2: raw MS output; sheet 3: list of detected alternative proteins; sheet 4: pie chart of corresponding altORFs localization.

Supplementary file 3: list of phosphopeptides.

Supplementary file 4: linker sequences separating adjacent zinc finger motifs.

Supplementary file 5: 260 alternative proteins detected by mass spectrometry in the

927 **interactome of 131 zinc finger proteins.** Sheet 1: MS identification parameters; sheet 2:
928 raw MS output; sheet 3: list of detected alternative proteins.
929
930 **Supplementary file 6: high-confidence list of predicted functional and co-operating**
931 **alternative proteins based on co-conservation and expression analyses.** Sheet 1: co-
932 conservation in mammals; sheet 2: co-conservation in vertebrates.

Figure 1

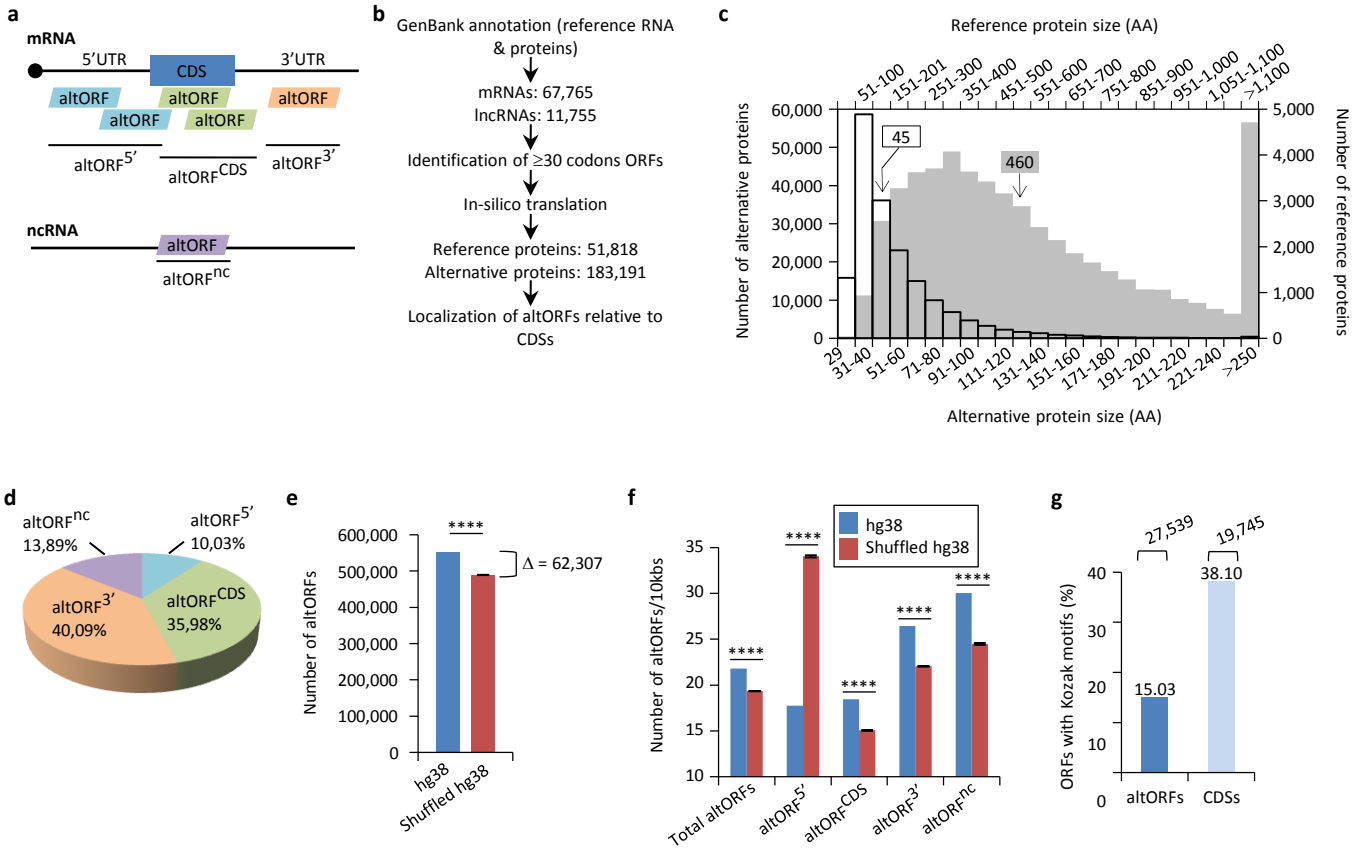


Figure 1. Annotation of human altORFs.
(**a**) AltORF nomenclature. AltORFs partially overlapping the CDS must be in a different reading frame. (**b**) Pipeline for the identification of altORFs. (**c**) Size distribution of alternative (empty bars, vertical and horizontal axes) and reference (grey bars, secondary horizontal and vertical axes) proteins. Arrows indicate the median size. The median alternative protein length is 45 amino acids (AA) compared to 460 for the reference proteins. (**d**) Distribution of altORFs in the human hg38 transcriptome. (**e, f**) Number of total altORFs (**e**) or number of altORFs/10kbs (**f**) in hg38 compared to shuffled hg38. Means and standard deviations for 100 replicates obtained by sequence shuffling are shown. Statistical significance was determined by using one sample t-test with two-tailed *p*-values. **** $P<0,0001$. (**g**) Percentage of altORFs with an optimal Kozak motif. The total number of altORFs with an optimal Kozak motif is also indicated at the top.

Figure 2

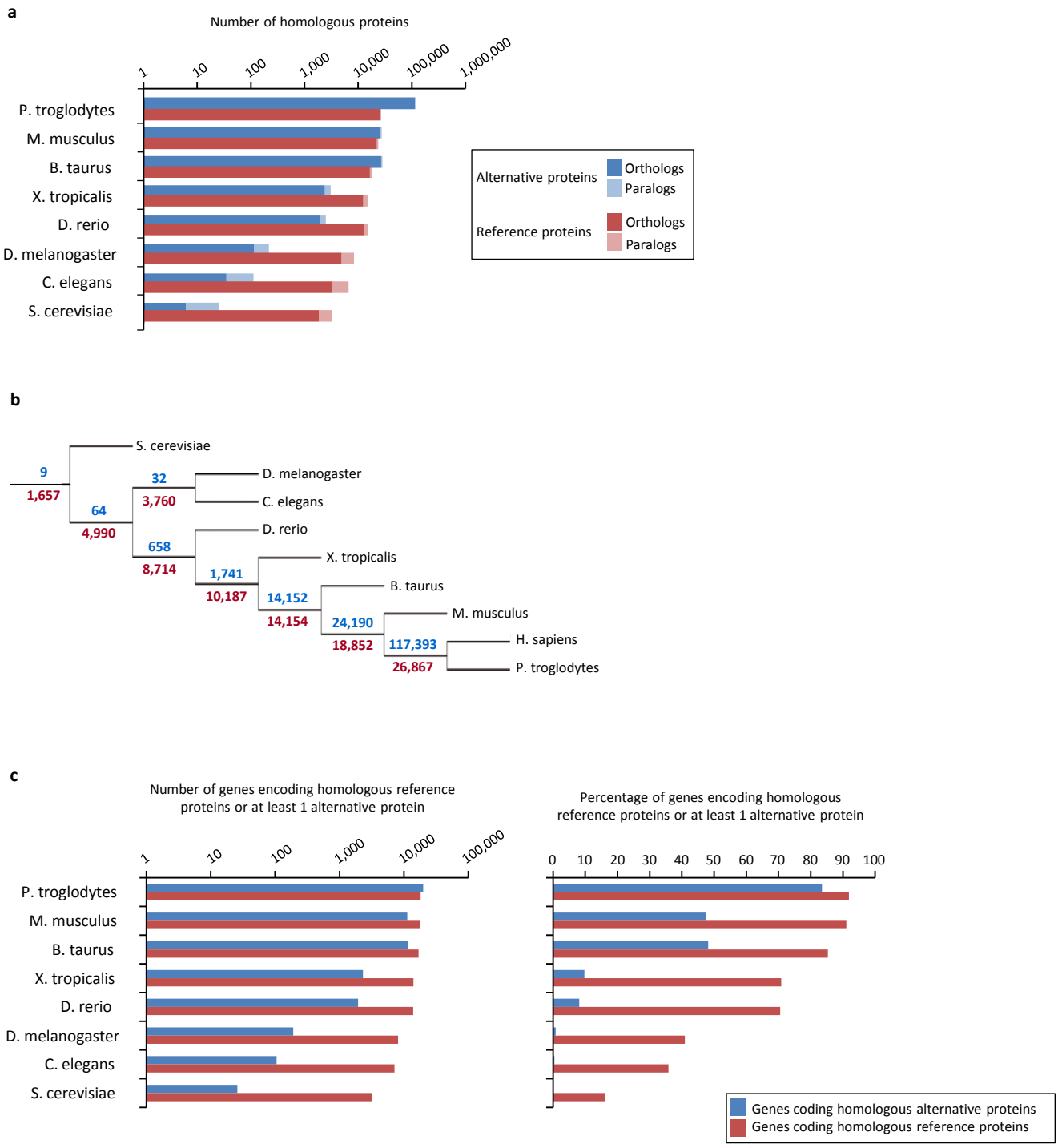


Figure 2: Conservation of alternative and reference proteins across different species. (**a**) Number of orthologous and paralogous alternative and reference proteins between *H. sapiens* and other species (pairwise study). (**b**) Phylogenetic tree: conservation of alternative (blue) and reference (red) proteins across various eukaryotic species. (**c**) Number and fraction of genes encoding homologous reference proteins or at least 1 homologous alternative protein.

Figure 3

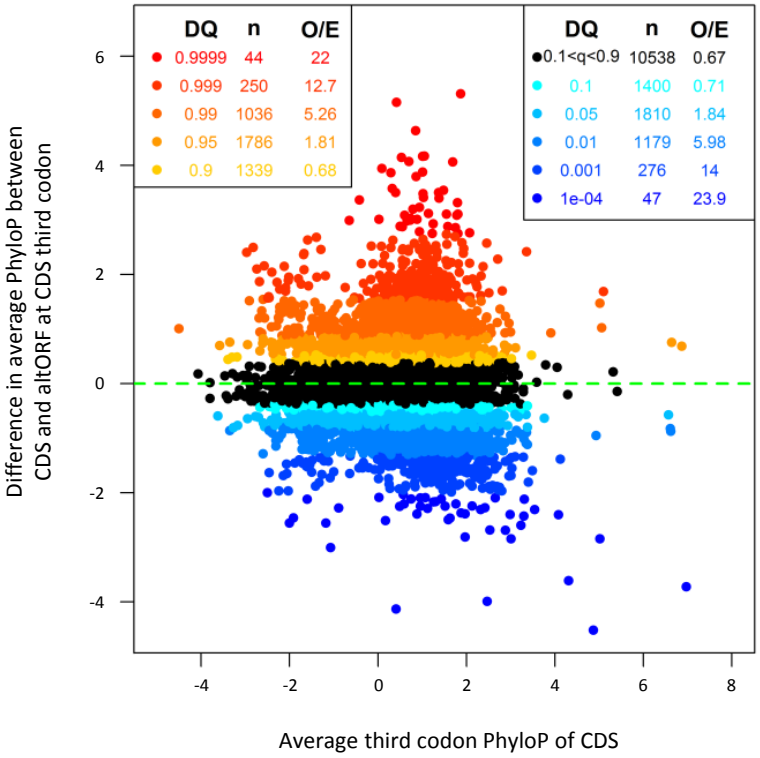


Figure 3: AltORFs completely nested within CDSs show more extreme PhyloP values (more conserved or faster evolving) than their CDSs. Differences between altORF and CDS PhyloP scores (altORF PhyloP – CDS PhyloP, *y*-axis) are plotted against PhyloPs for their respective CDSs (*x*-axis). The plot contains all 20,814 CDSs containing at least one fully nested altORF, paired with one of its altORFs selected at random (to avoid problems with statistical non-independence). PhyloPs for both altORFs and CDSs are based on 3rd codons in the CDS reading frame, calculated across 100 vertebrate species. We compared these differences to those generated based on five random regions in CDSs with a similar length as altORFs. Expected quantiles of the differences ("DQ" columns) were identified and compared to the observed differences. We show the absolute numbers ("n") and observed-to-expected ratios ("O/E") for each quantile. There are clearly substantial over-representations of extreme values (red signalling conservation DQ≥0.95, and blue signalling accelerated evolution DQ≤0.05) with 6,428 of 19,705 altORFs (36.2%). A random distribution would have implied a total of 10% (or 1,970) of altORFs in the extreme values. This suggests that 26.2% (36.2%-10%) of altORFs (or 4,458) undergo specific selection different from random regions in their CDSs with a similar length distribution.

Figure 4

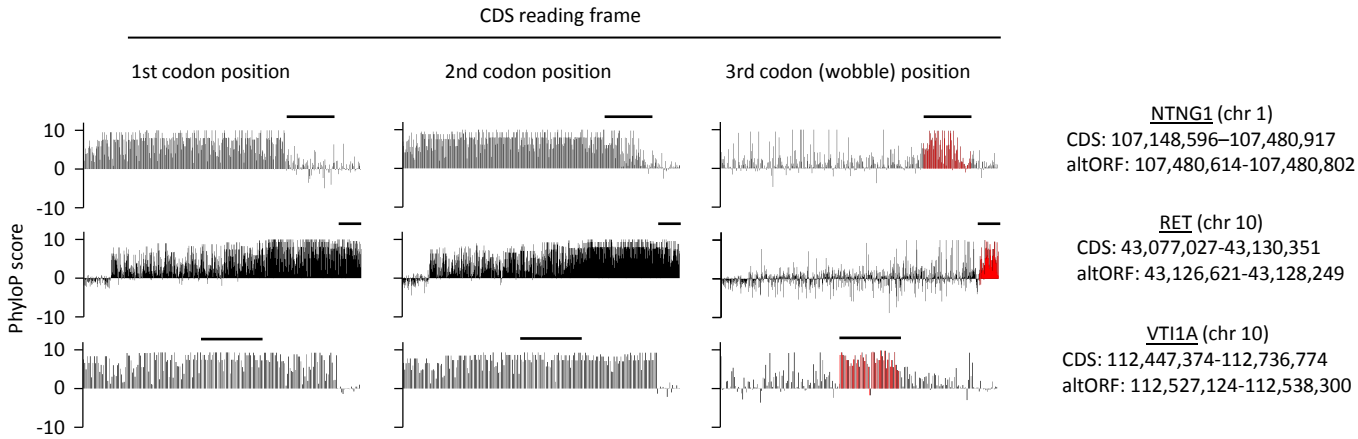


Figure 4: First, second, and third codon nucleotide PhyloP scores for 100 vertebrate species for the CDSs of the NTNG1, RET and VTI1A genes. Chromosomal coordinates for the different CDSs and altORFs are indicated on the right. The regions highlighted in red indicate the presence of an altORF characterized by a region with elevated PhyloP scores for wobble nucleotides. The region of the altORF is indicated by a black bar above each graph.

Figure 5

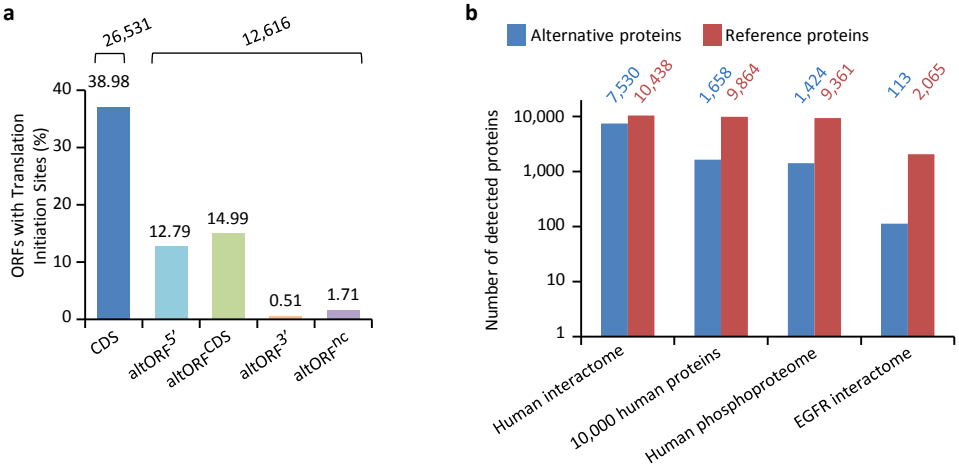


Figure 5. Expression of human altORFs.
(**a**) Percentage of CDSs and altORFs with detected TISs by ribosomal profiling and footprinting of human cells[23]. The total number of CDSs and altORFs with a detected TIS is indicated at the top. (**b**) Alternative and reference proteins detected in three large proteomic datasets: human interactome[28], 10,000 human proteins[31], human phosphoproteome[30], EGFR interactome[29]. Numbers are indicates above each column.

Figure 6

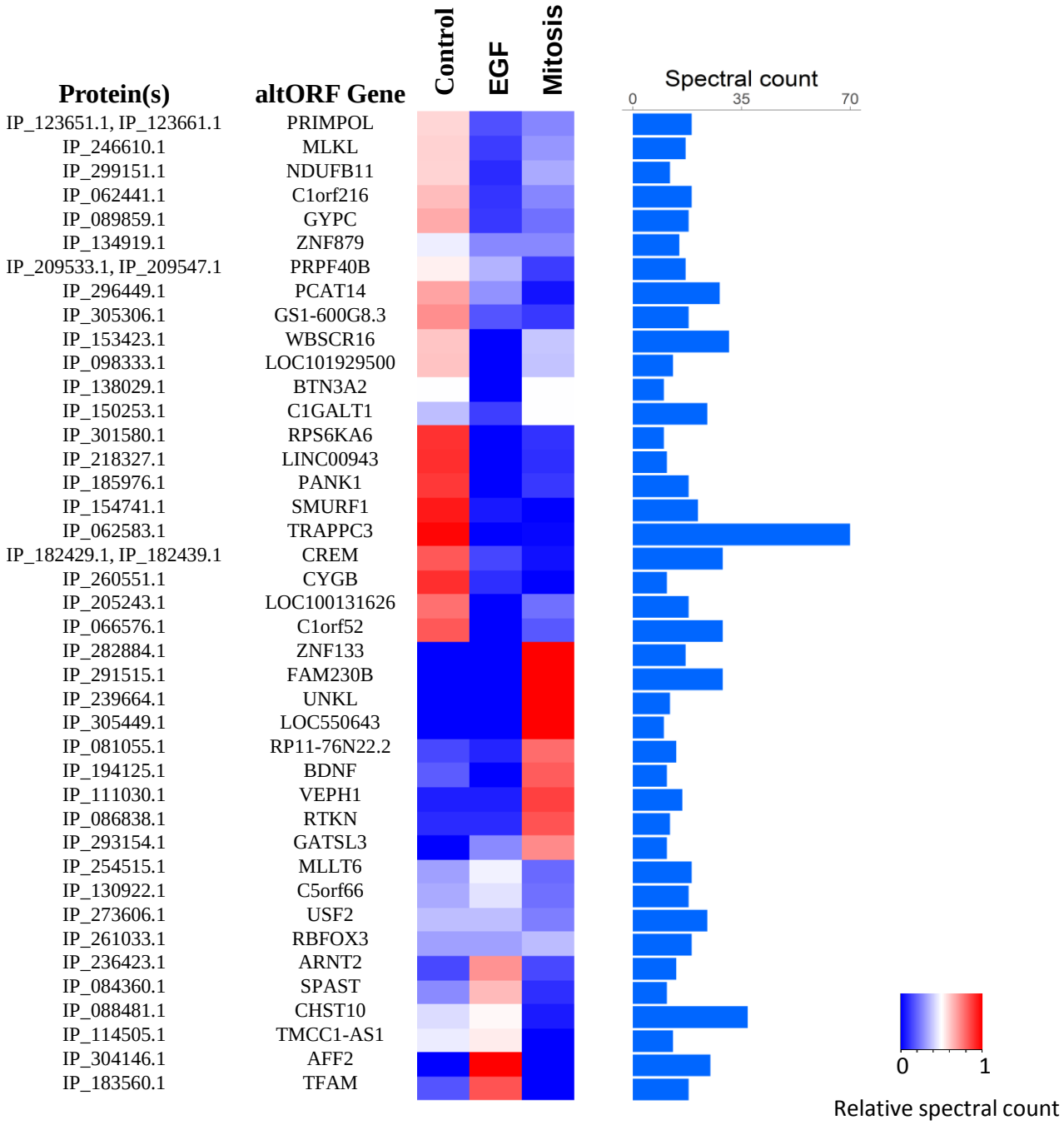


Figure 6: The alternative phosphoproteome in mitosis and EGF-treated cells. Heatmap showing relative levels of spectral counts for phosphorylated peptides following the indicated treatment[29]. For each condition, heatmap colors show the percentage of spectral count on total MS/MS phosphopeptide spectra. Blue bars on the right represent the number of MS/MS spectra; only proteins with spectral counts covering a range between 70 and 10 are shown.

Figure 7

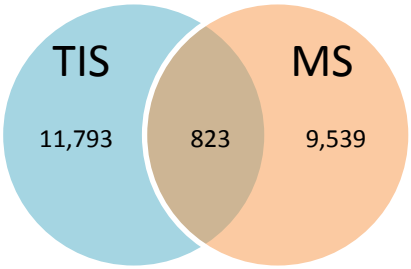


Figure 7: Number of alternative proteins detected by ribosome profiling and mass spectrometry.
The expression of 823 alternative proteins was detected by both ribosome profiling (translation initiation sites, TIS) and mass spectrometry (MS).

Figure 8

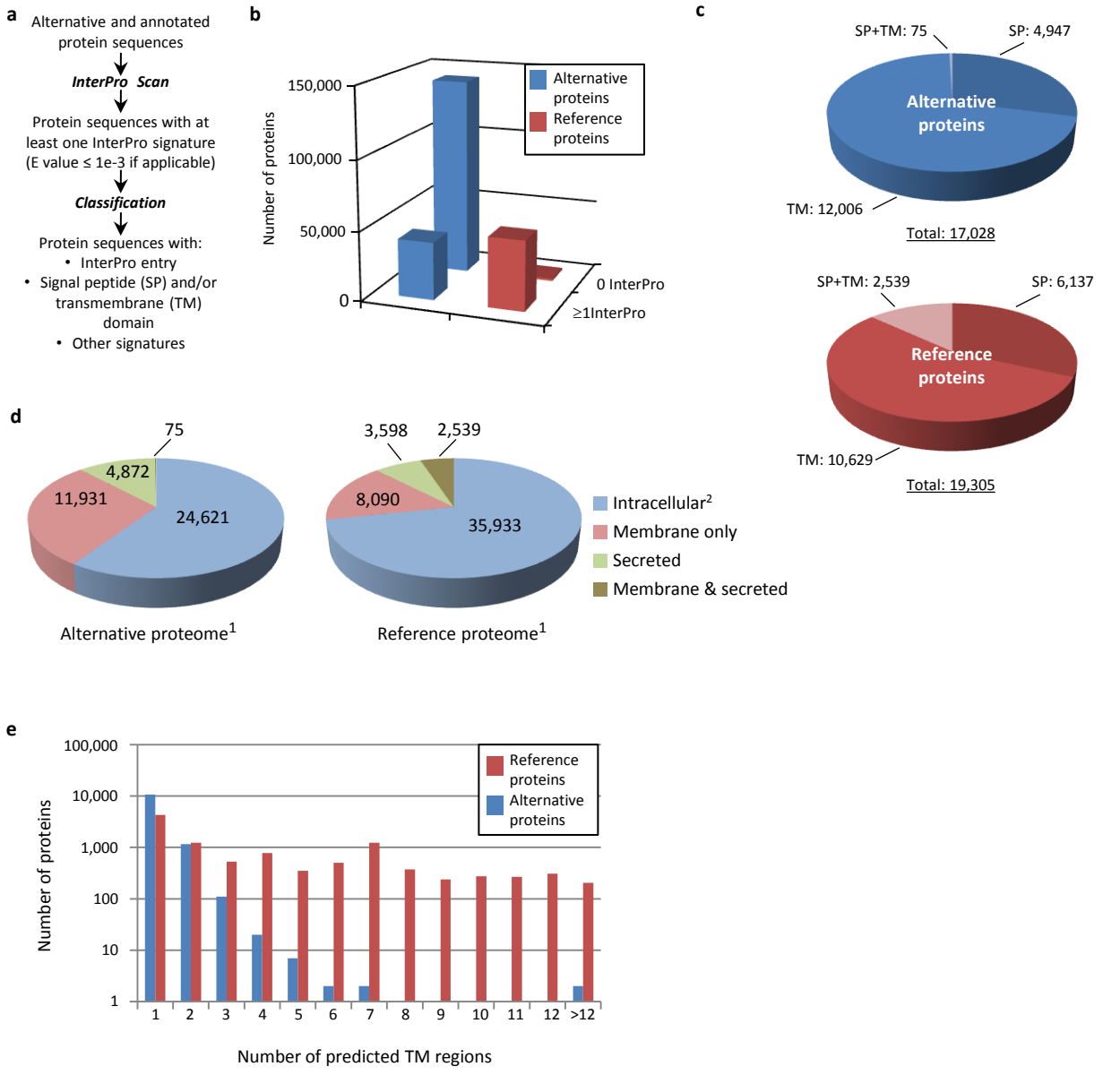


Figure 8: Human alternative proteome sequence analysis and classification using InterProScan.

(**a**) InterPro annotation pipeline. (**b**) Alternative and reference proteins with InterPro signatures. (**c**) Number of alternative and reference proteins with transmembrane domains (TM), signal peptides (S) and both TM and SP. (**d**) Number of all alternative and reference proteins predicted to be intracellular, membrane, secreted and membrane-spanning and secreted. [1]Proteins with at least one InterPro signature; [2]Proteins with no predicted signal peptide or transmembrane features. (**e**) Number of predicted TM regions for alternative and reference proteins.

Figure 9

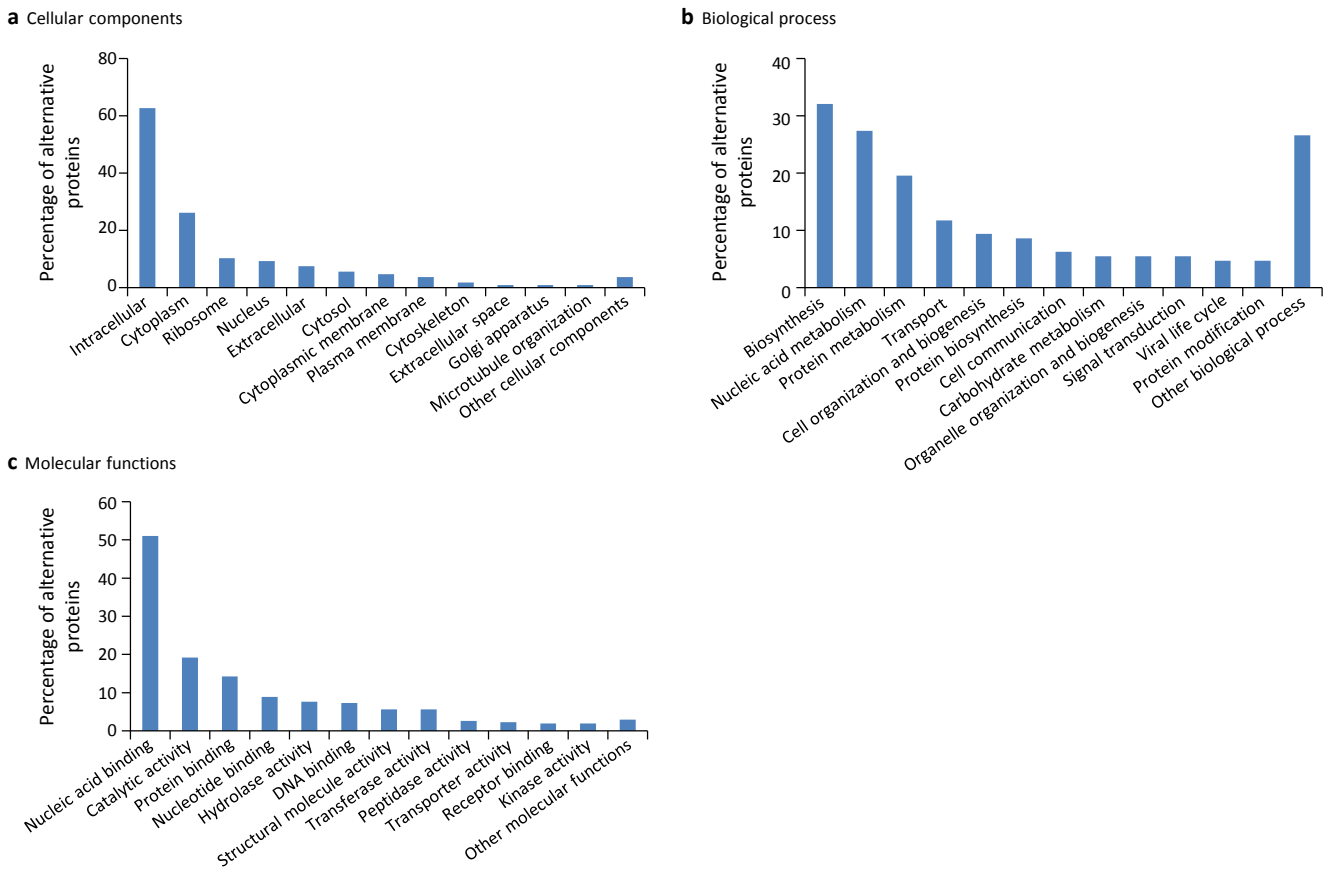


Figure 9: Gene ontology (GO) annotations for human alternative proteins.
GO terms assigned to InterPro entries are grouped into 13 categories for each of the three ontologies. (**a**) 34 GO terms were categorized into cellular component for 107 alternative proteins. (**b**) 64 GO terms were categorized into biological process for 128 alternative proteins. (**c**) 94 GO terms were categorized into molecular function for 302 alternative proteins. The majority of alternative proteins with GO terms are predicted to be intracellular, to function in nucleic acid-binding, catalytic activity and protein binding and to be involved in biosynthesis and nucleic acid metabolism processes.

Figure 10

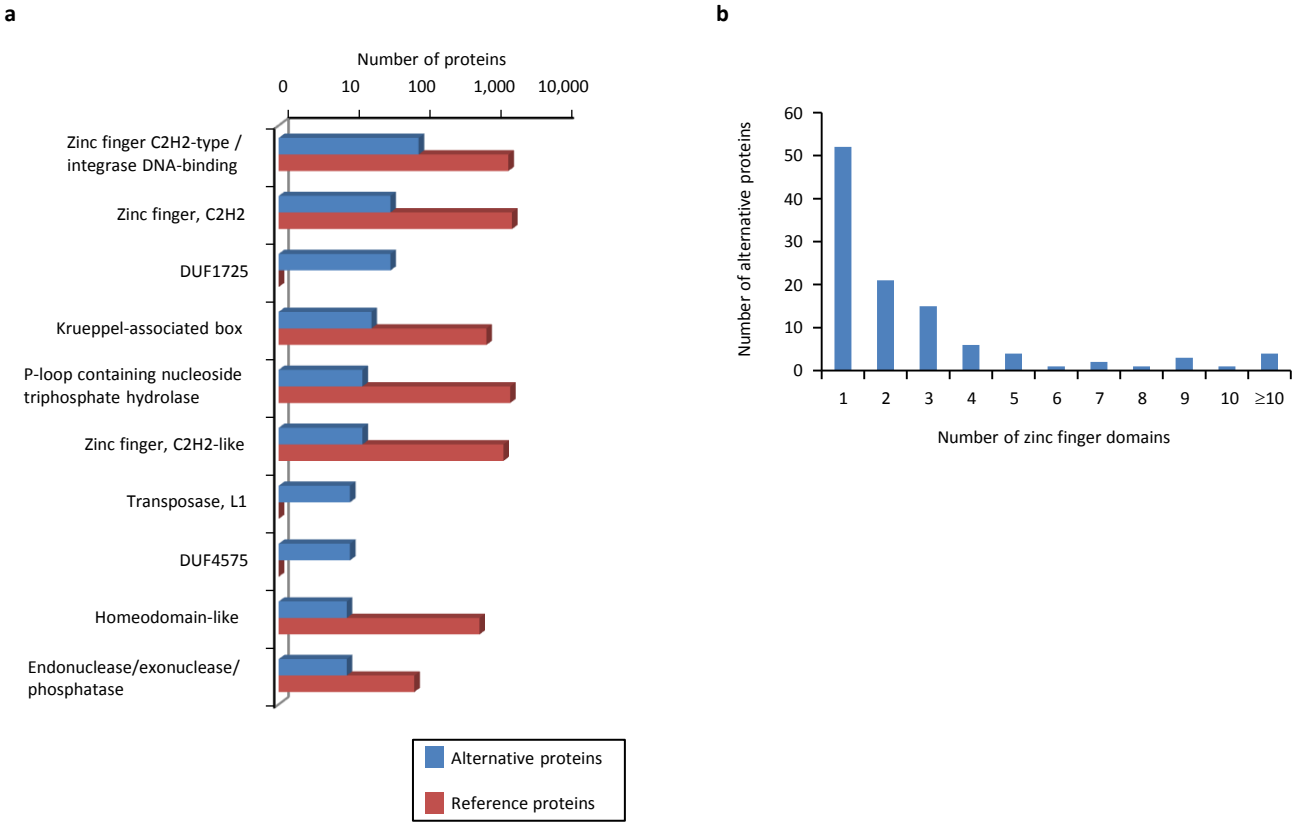


Figure 10: Main InterPro entries in human alternative proteins. (**a**) The top 10 InterPro families in the human alternative proteome. (**b**) A total of 110 alternative proteins have between 1 and 23 zinc finger domains.

Figure 11

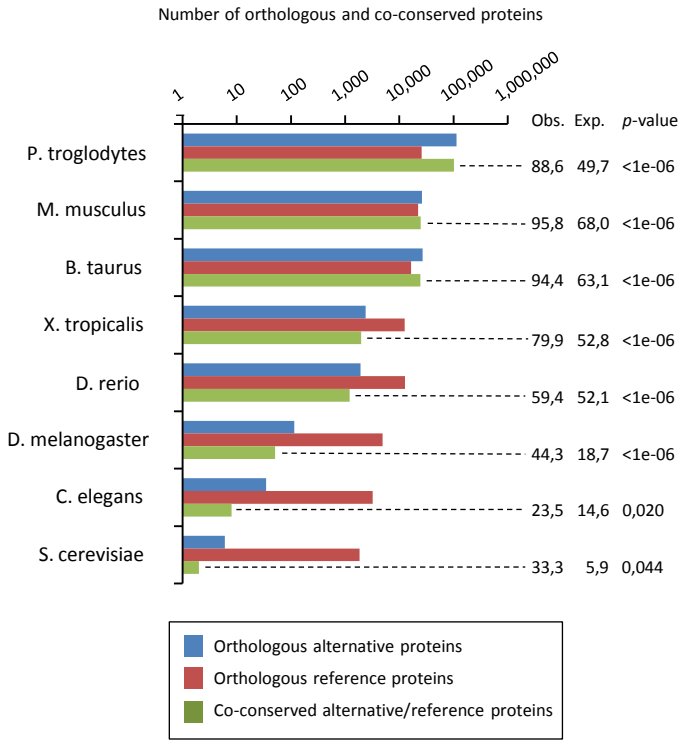


Figure 11: Number of orthologous and co-conserved alternative and reference proteins between *H. sapiens* and other species (pairwise). For the co-conservation analyses, the percentage of observed (Obs.), expected (Exp.) and corresponding *p*-values is indicated on the right (see Table 4 for details).

Figure 12

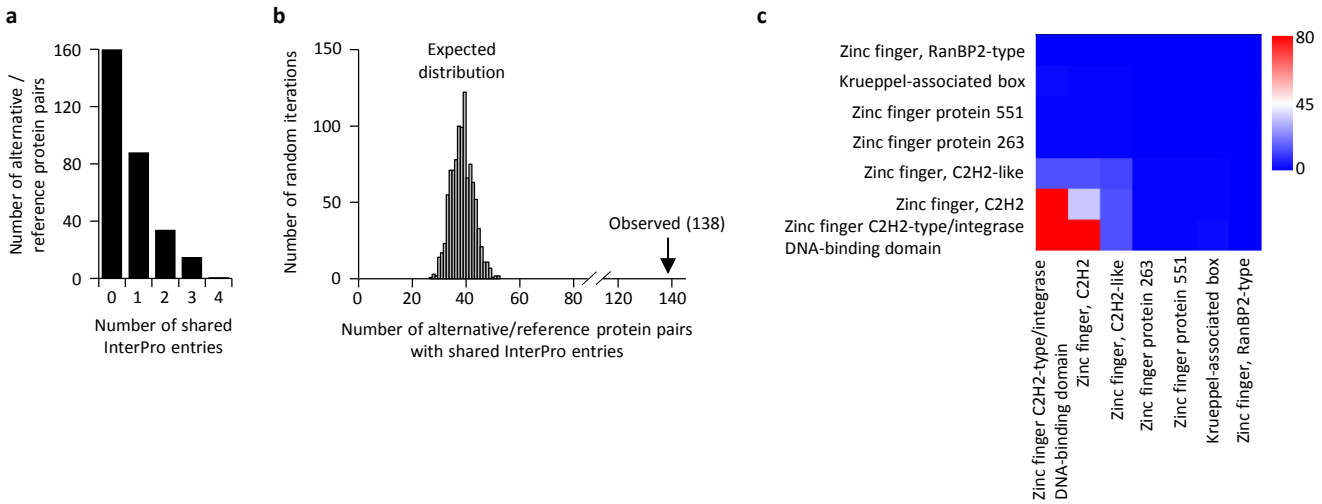


Figure 12. Reference and alternative proteins share functional domains.
(**a**) Distribution of the number of shared InterPro entries between alternative and reference proteins coded by the same transcripts. 138 pairs of alternative and reference proteins share between 1 and 4 protein domains (InterPro entries). Only alternative/reference protein pairs that have at least one domain are considered (n = 298). (**b**) The number of reference/alternative protein pairs that share domains (n = 138) is higher than expected by chance alone. The distribution of expected pairs sharing domains and the observed number are shown. (**c**) Matrix of co-occurrence of domains related to zinc fingers. The entries correspond to the number of times entries co-occur in reference and alternative proteins. The full matrix is available in figure 12-figure supplement 1.

Figure 13

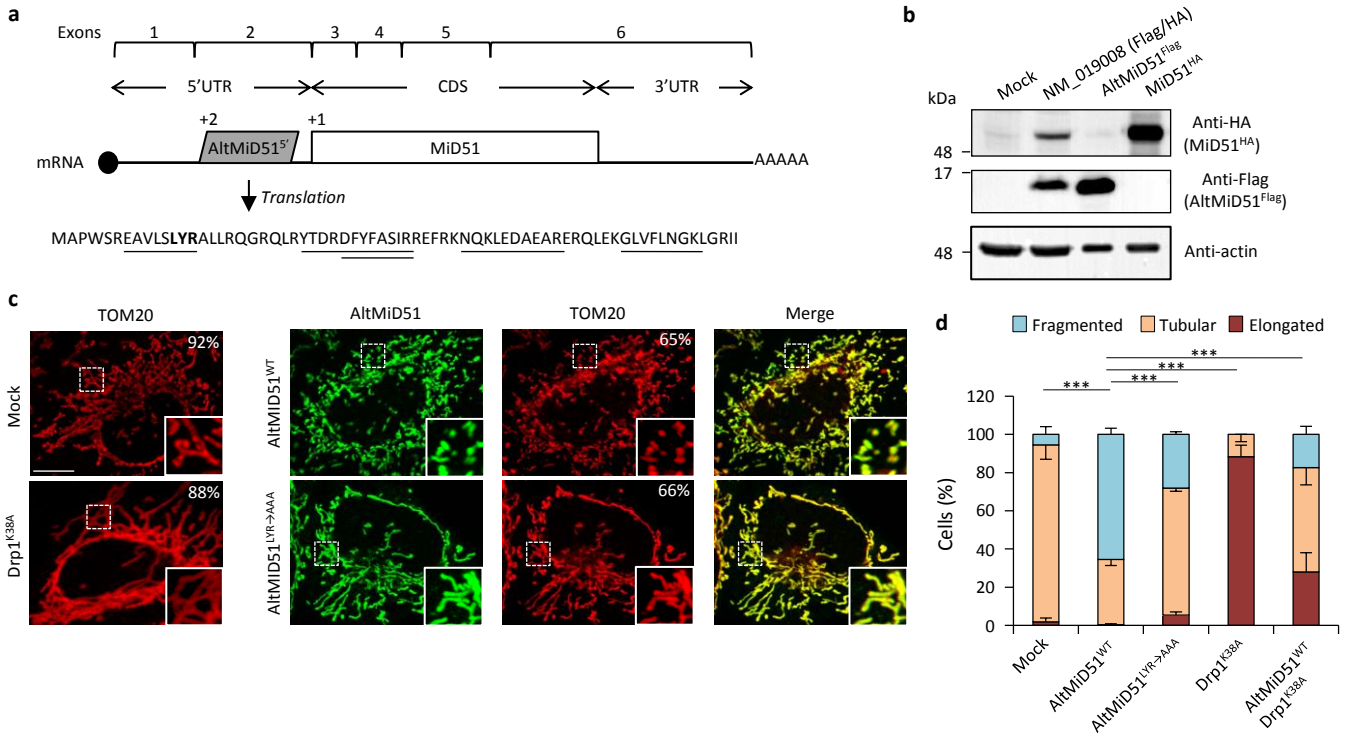


Figure 13. AltMiD51$^{5'}$ expression induces mitochondrial fission.
(**a**) AltMiD51$^{5'}$ coding sequence is located in exon 2 or the *MiD51/Mief1/SMCR7L* gene and in the 5'UTR of the canonical mRNA (RefSeq NM_019008). +2 and +1 indicate reading frames. AltMiD51 amino acid sequence is shown with the LYR tripeptide shown in bold. Underlined peptides were detected by MS. (**b**) Human HeLa cells transfected with empty vector (mock), a cDNA corresponding to the canonical MiD51 transcript with a Flag tag in frame with altMiD51 and an HA tag in frame with MiD51, altMiD51^{Flag} cDNA or MiD51^{HA} cDNA were lysed and analyzed by western blot with antibodies against Flag, HA or actin, as indicated. (**c**) Confocal microscopy of mock-transfected cells, cells transfected with altMiD51^{WT}, altMiD51$^{LYR\rightarrow AAA}$ or Drp1^{K38A} immunostained with anti-TOM20 (red channel) and anti-Flag (green channel) monoclonal antibodies. In each image, boxed areas are shown at higher magnification in the bottom right corner. % of cells with the most frequent morphology is indicated: mock (tubular), altMiD51^{WT} (fragmented), altMiD51(LYR→AAA) (tubular), Drp1(K38A) (elongated). Scale bar, 10 mm. (**d**) Bar graphs show mitochondrial morphologies in HeLa cells. Means of three independent experiments per condition are shown. ***p<0.0005 (Fisher's exact test) for the three morphologies between altMiD51(WT) and the other experimental conditions.

Figure 14

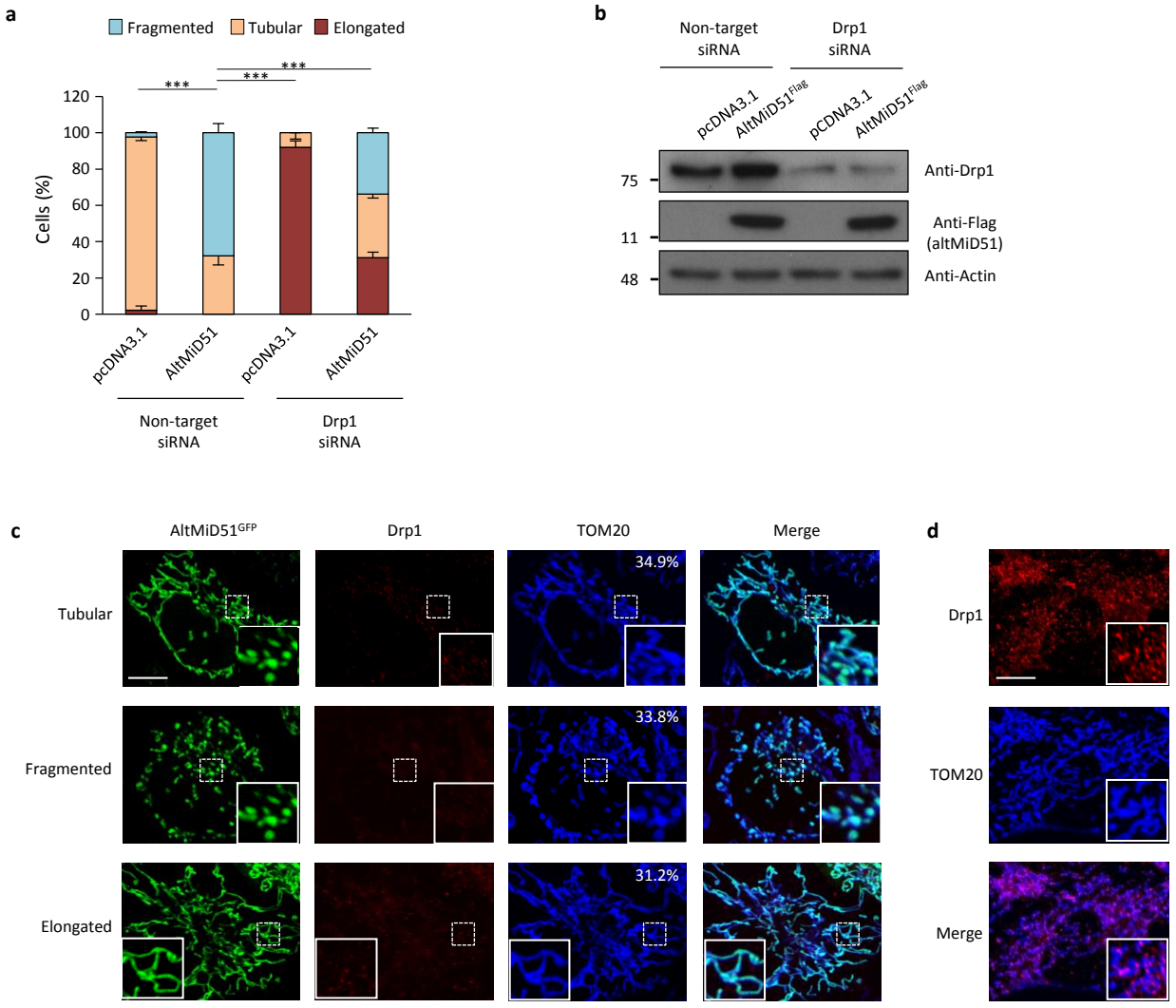


Figure 14: AltMiD51-induced mitochondrial fragmentation is dependent on Drp1. (**a**) Bar graphs show mitochondrial morphologies in HeLa cells treated with non-target or Drp1 siRNAs. Cells were mock-transfected (pcDNA3.1) or transfected with altMiD51^{Flag}. Means of three independent experiments per condition are shown. ***p<0.0005 (Fisher's exact test) for the three morphologies between altMiD51 and the other experimental conditions. (**b**) HeLa cells treated with non-target or Drp1 siRNA were transfected with empty vector (pcDNA3.1) or altMiD51^{Flag}, as indicated. Proteins were extracted and analyzed by western blot with antibodies against the Flag tag (altMiD51), Drp1 or actin, as indicated. Molecular weight markers are shown on the left (kDa). (**c**) Confocal microscopy of Drp1 knockdown cells transfected with altMiD51^{GFP} immunostained with anti-TOM20 (blue channel) and anti-Drp1 (red channel) monoclonal antibodies. In each image, boxed areas are shown at higher magnification in the bottom right corner. % of cells with the indicated morphology is indicated on the TOM20 panels. Scale bar, 10 μm. (**d**) Control Drp1 immunostaining in HeLa cells treated with a non-target siRNA. For (**c**) and (**d**), laser parameters for Drp1 and TOM20 immunostaining were identical.

Figure 13

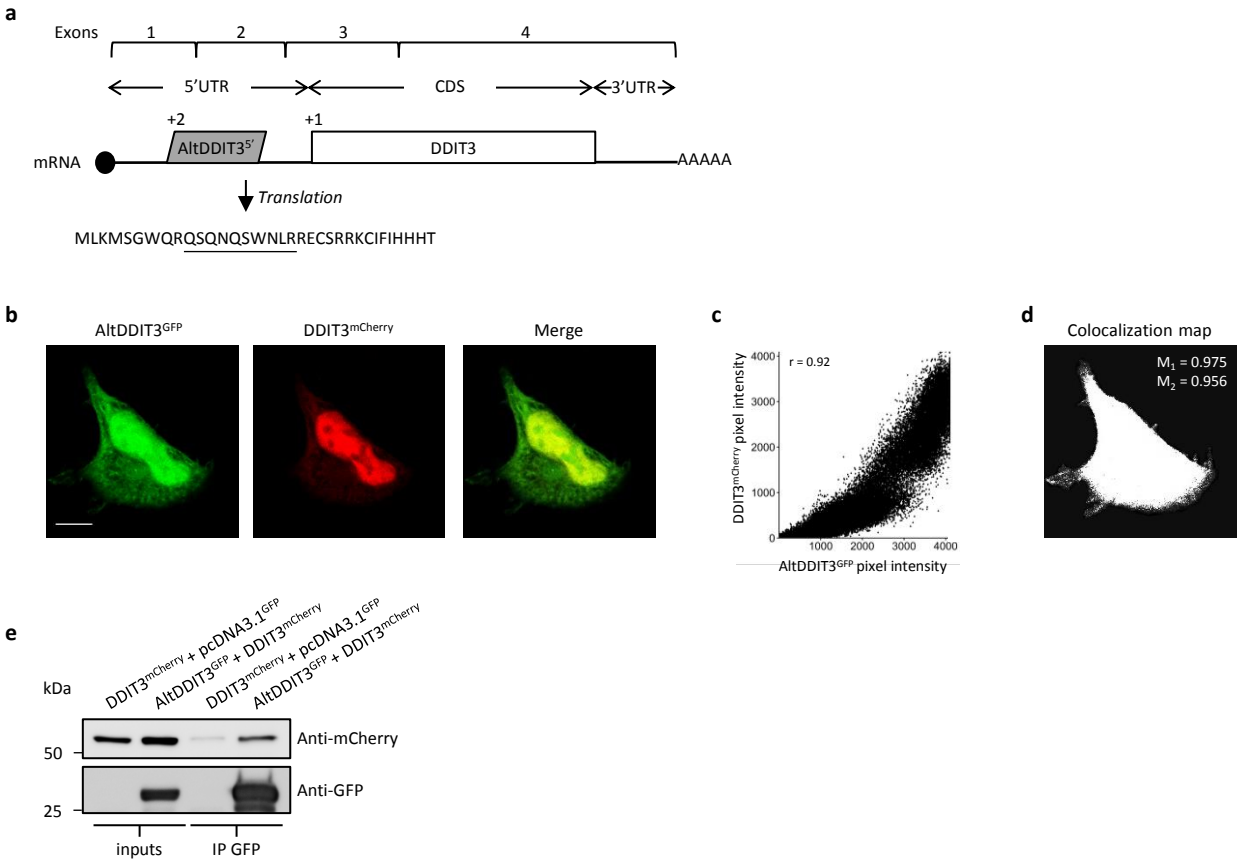


Figure 15. AltDDIT3$^{5'}$ co-localizes and interacts with DDIT3.
(**a**) AltDDIT3$^{5'}$ coding sequence is located in exons 1 and 2 or the *DDIT3/CHOP/GADD153* gene and in the 5'UTR of the canonical mRNA (RefSeq NM_001195053). +2 and +1 indicate reading frames. AltDDIT3 amino acid sequence is shown with the underlined peptide detected by MS. (**b**) Confocal microscopy analyses of HeLa cells co-transfected with altDDIT3^{eGFP} (green channel) and DDIT3^{mCherry} (red channel). Scale bar, 10 µm. (**c, d**) Colocalization analysis of the images shown in (**c**) performed using the JACoP plugin (Just Another Co-localization Plugin) implemented in Image J software. (**c**) Scatterplot representing 50 % of green and red pixel intensities showing that altDDIT3^{GFP} and DDIT3^{mCherry} signal highly correlate (with Pearson correlation coefficient of 0.92 (p-value < 0.0001)). (**d**) Binary version of the image shown in (c) after Costes' automatic threshold. White pixels represent colocalization events (p-value < 0.001, based on 1000 rounds of Costes' randomization colocalization analysis). The associated Manders Correlation Coefficient, M_1 and M_2, are shown in the right upper corner. M_1 is the proportion of altDDIT3^{GFP} signal overlapping DDIT3^{mCherry} signal and M_2 is the proportion of DDIT3^{mCherry} signal overlapping altDDIT3^{GFP}. (**e**) Representative immunoblot of co-immunoprecipitation with GFP-Trap agarose beads performed on HeLa lysates co-expressing DDIT3^{mcherry} and altDDIT3^{GFP} or DDIT3^{mcherry} with pcDNA3.1^{GFP} empty vector (n = 2).

Table 1: AltORFs and alternative protein annotations in different organisms

Genomes	Features					
	Transcripts		Current annotations		Annotations of alternative protein coding sequences	
	mRNAs	Others[1]	CDSs	Proteins	altORFs	Alternative proteins
H. sapiens GRCh38 RefSeq GCF_000001405.26	67,765	11,755	68,066	51,818	539,895	183,191
P. troglodytes 2.1.4 RefSeq GCF_000001515.6	55,034	7,527	55,243	41,774	416,515	161,663
M. musculus GRCm38p2, RefSeq GCF_000001635.22	73,450	18,886	73,551	53,573	642,203	215,472
B. Taurus UMD3.1.86	22,089	838	22,089	21,915	79,906	73,603
X. tropicalis Ensembl JGI_4.2	28,462	4,644	28,462	22,614	141,894	69,917
D rerio Ensembl ZV10.84	44,198	8,196	44,198	41,460	214,628	150,510
D. melanogaster RefSeq GCA_000705575.1	30,255	3,474	30,715	20,995	174,771	71,705
C. elegans WBcel235, RefSeq GCF_000002985.6	28,653	25,256	26,458	25,750	131,830	45,603
S. cerevisiae YJM993_v1, RefSeq GCA_000662435.1	5,471	1,463	5,463	5,423	12,401	9,492

[1]Other transcripts include miRNAs, rRNAs, ncRNAs, snRNAs, snoRNAs, tRNAs. [2]Annotated retained-intron and processed transcripts were classified as mRNAs.

Table 2: alternative zinc finger proteins detected by mass spectrometry (MS) and ribosome profiling (RP)

Alternative protein accession	Detection method[1]	Gene	Amino acid sequence	AltORF localization
IP_238718.1	MS	RP11	MLVEVACSSCRSLLHKGAGASEDGAALEPAHTGGKENGATT	nc
IP_278905.1	MS and RP	ZNF761	MSVARPLVGSHILYAIIDFILERNLISVMSVARTLVRSHPLYAT IDFILERNLTSVMSVARPLVRSQTLHAIVDFILEKNKCNECGE VFNQQAHLAGHHRIHTGEKP	CDS
IP_278681.1	MS	ZNF468	MNVARFLIKKQPLHITIDFILERNLTNGRNVTKVFSCKSNLKT HKKIHIEEKPYRGKVCDKVFAYNAYLAKHTRIHTGEKLIISVM SVARPLVKIHTL	3’
IP_106493.1	MS	ZNF717	MWKNLSSQVIPHHTPENSHGEKPYGCNECGKTFCQKSYLIIH QRTHTGEKPYECNECGKSFHQKANLQKHQGIHTGEKPYECS KCGKTLSEVSPHCTS	CDS
IP_278745.1	MS and RP	ZNF816	MSVARPSVRNHPFNAIIYFTLERNLTNVKNVTMFTFADHTLK DIGRFILERDHTNVRFVTRFSGVIHTLQNIREFILERNHTSVINV AGVSVGSHPFNTIIHFTLERNLTHVMNVARFLVEEKTLHVIID FMLERNLTNVKNVTKFSVADHTLKDIGEFILGKNHTNVRFVT RLSGVIHALQTIREFILERNLTSVINVRRFLIKKESLHNIREFILE RNLTSVMNVARFLIKKQALQNIREFILQRNLTSVMSVAKPLL DSQHLFTIKQSMGVGKLYKCNDCHKVFSNATTIANHYRIHIE ERSTSVINVANFSDVIHNL	CDS
IP_138289.1	MS	ZSCAN31	MNIGGATLERNPINVRSVGKPSVPAMASLDTEESTQGKNHM NAKCVGRLSSSAHALFSIRGYTLERSAISVVSVAKPSFRMQGF SSISESTLVRNPISAVSAVNSLVSGHFLRNIRKSTLERDHKGDE FGKAFSHHCNLIRHFRIHTVPAELD	CDS
IP_278564.1	MS	ZNF808	MIVTKSSVTLQQLQIIGESMMKRNLLSVINVACFSDIVHTLQFI GNLILERNLTNVMIEARSSVKLHPMQNRRIHTGEKPHKCDDC GKAFTSHSHLVGHQRIHTGQKSCKCHQCGKVFSPRSLLAEHE KIHF	3’
IP_275012.1	MS	ZNF780A	MKPCECTECGKTFSCSSNIVQHVKIHTGEKRYNVRNMGKHLL WMISCLNIRKFRIVRNFVTIRSVDKPSLCTKNLLNTRELILMRN LVNIKECVKNFHHGLGFAQLLSIHTSEKSLSVRNVGRFIATLN TLEFGEDNSCEKVFE	3’
IP_204754.1	RP	ZFP91-CNTF	MPGETEEPRPPEQQDQEGGEAAKAAPEEPQQRPPEAVAAAPA GTTSSRVLRGGRDRGRAAAAAAAAAVSRRRKAEYPRRRRSS PSARPPDVPGQQPQAAKSPSPVQGKKSPRLLCIEKVTTDKDPK EEKEEEDDSALPQEVSIAASRPSRGWRSSRTSVSRHRDTENTR SSRSKTGSLQLICKSEPNTDQLDYDVGEEHQSPGGISSEEEEEE EEEMLISEEEIPFKDDPRDETYKPHLERETPKPRRKSGKVKEE KEKKEIKVEVEVEVKEEENEIREDEEPPRKRGRRRKDDKSPRL PKRRKKPPIQYVRCEMEGCGTVLAHPRYLQHHIKYQHLLKK KYVCPHPSCGRLFRLQKQLLRHAKHHTDQRDYICEYCARAF KSSHNLAVHRMIHTGEKPLQCEICGFTCRQKASLNWHMKKH DADSFYQFSCNICGKKFEKKDSVVAHKAKSHPEVLIAEALAA NAGALITSTDILGTNPESLTQPSDGQGLPLLPEPLGNSTSGECL LLEAEGMSKSYCSGTERSIHR	nc
IP_098649.1	RP	INO80B-WBP1	MSKLWRRGSTSGAMEAPEPGEALELSLAGAHGHGVHKKKH KKHKKKHKKKHHQEEDAGPTQPSPAKPQLKLKIKLGGQVLG TKSVPTFTVIPEGPRSPSPLMVVDNEEEPMEGVPLEQYRAWL DEDSNLSPSPLRDLSGGLGGQEEEEEQRWLDALEKGELDDNG DLKKEINERLLTARQRALLQKARSQPSPMLPLPVAEGCPPPAL TEEMLLKREERARKRRLQAARRAEEHKNQTIERLTKTAATSG RGGRGGARGERRGGRAAAPAPMVRYCSGAQGSTLSFPPGVP APTAVSQRPSPSGPPPRCSVPGCPHPRRYACSRTGQALCSLQC YRINLQMRLGGPEGPGSPLLATFESCAQE	nc
IP_115174.1	RP	ZNF721	MYIGEFILERNPTHVENVAKPLDSLQIFMRIRKFILERNPTRVE TVAKPLDSLQIFMHIRKFILEIKPYKCKECGKAFKSYYSILKHK	CDS

			RTHTRGMSYEGDECRGL	
IP_275016.1	RP	ZNF780A	MNVRSVGKALIVVHTLFSIRKFIPMRNLLYVGNVRWPLDIIANLLNILEFILVTSHLNVKTVGRPSIVAQALFNIRVFTLVRSPMNVRSVGRLLDFTYNFPNIRKLTQVKNHLNVRNVGNSFVVVQILINIEVFILERNPLNVRNVGKPFDFICTLFDIRNCILVRNPLNVRSVGKPFDFICNLFDIRNCILVRNPLNVRNVERFLVFPPSLIAIRTFTQVRRHLECKECGKSFNRVSNHVQHQSIRAGVKPCECKGCGKGFICGSNVIQHQKIHSSEKLFVCKEWRTTFRYHYHLFNITKFTLVKNPLNVKNVERPSVF	CDS or 3’
IP_278870.1	RP	ZNF845	MNVARFLIEKQNLHVIIEFILERNIRNMKNVTKFTVVNQVLKDRRIHTGEKAYKCKSL	CDS
IP_278888.1	RP	ZNF765	MSVARPSAGRHPLHTIIDFILDRNLTNVKIVMKLSVSNQTLKDIGEFILERNYTCNECGKTFNQELTLTCHRRLHSGEKPYKYEELDKAYNFKSNLEIHQKIRTEENLTSVMSVARP	CDS
IP_278918.1	RP	ZNF813	MNVARVLIGKHTLHVIIDFILERNLTSVMNVARFLIEKHTLHIIIDFILEINLTSVMNVARFLIKKHTLHVTIDFILERNLTSVMNVARFLIKKQTLHVIIDFILERNLTSLMSVAKLLIEKQSLHIIIQFILERNKCNECGKTFCHNSVLVIHKNSYWRETSVMNVAKFLINKHTFHVIIDFIVERNLRNVKHVTKFTVANRASKDRRIHTGEKAYKGEEYHRVFSHKSNLERHKINHTAEKP	CDS
IP_280349.1	RP	ZNF587	MNAVNVGNHFFPALRFMFIKEFILDKSLISAVNVENPFLNVPVSLNTGEFTLEKGLMNAPNVEKHFSEALPSFIIRVHTGERPYECSEYGKSFAEASRLVKHRRVHTGERPYECCQCGKHQNVCCPRS	CDS
IP_280385.1	RP	ZNF417	MNAMNVGNHFFPALRFMFIKEFILDKSLISAVNVENPLLNVPVSLNTGEFTLEKGLMNVPNVEKHFSEALPSFIIRVHTGERPYECSEYGKSFAETSRLIKHRRVHTGERPYECCQSGKHQNVCSPWS	CDS

[1]MS, mass spectrometry; RP, ribosome profiling.

Table 3: Examples of proteins encoded in the same gene and functionally interacting

Gene	Polypeptides[1]	Reference
CDKN2A, INK4	Cyclin-dependent kinase inhibitor 2A or p16-INK4 (P42771), and p19ARF (Q8N726)	(47)
GNAS, XLalphas	Guanine nucleotide-binding protein G(s) subunit alpha isoforms XLαs (Q5JWF2) and Alex (P84996)	(48)
ATXN1	Ataxin-1 (P54253) and altAtaxin-1	(49)
Adora2A	A2A adenosine receptor (P30543) and uORF5	(50)
AGTR1	Angiotensin type 1a receptor (P25095) and PEP7	(51)

[1]The UniProtKB accession is indicated when available.

Table 4: orthology and co-conservation assessment of alternative-reference protein pairs between H. sapiens and other species

	A	B	C	D	E	F	G	H	I	J
						Observed	Mean expected		Max expected	
	Orthologous altProts (of 183,191 total)	Orthologous refProts	Co-conserved altProt-refProt pairs	Non-orthologous altProts	Non-orthologous altProts paired with orthologous refProts	Co-conservation (C/A)	% orthologous refProts (B/51,819)	% non-orthologous altProts paired with an orthologous refProt (E/D)	Max % of 1 million binomial simulations, p=max(G, H), n=A	Inferred *p*-value
P. troglodytes	113,687	25,755	100,839	69,504	30,772	88.69	49.70	44.27	50.39	<1e-06
M. musculus	25,930	22,304	24,862	157,261	106,987	95.88	43.04	68.031	69.39	<1e-06
B. taurus	25,868	16,887	24,426	157,323	99,369	94.42	32.58	63.16	64.67	<1e-06
X. tropicalis	2,470	12,458	1,974	180,721	95,499	79.91	24.04	52.84	57.81	<1e-06
D. rerio	2,023	12,791	1,203	181,168	94,426	59.46	24.68	52.12	57.29	<1e-06
D. melanogaster	115	4,881	51	183,076	34,352	44.34	9.41	18.76	38.26	<1e-06
C. elegans	34	3,954	8	183,157	26,839	23.52	7.63	14.65	50.00	0.02
S. cerevisiae	6	1,854	2	183,185	10,935	33.33	3.57	5.96	83.33	0.04

In order to compare the observed co-conservation to expected co-conservation, we used the more conservative of two expected values: either the percentage of all refProts (called here reference proteins) that were defined as orthologous (column G), or the percentage of non-orthologous altProts (called here alternative proteins) that were paired with an orthologous refProt. Both of these methods are themselves conservative, as they do not account for the conservation of the pairing. The larger of these values for each species was then used to generate 1 million random binomial distributions with n=# of orthologous altProts; the maximum of these percentages is reported in column I.

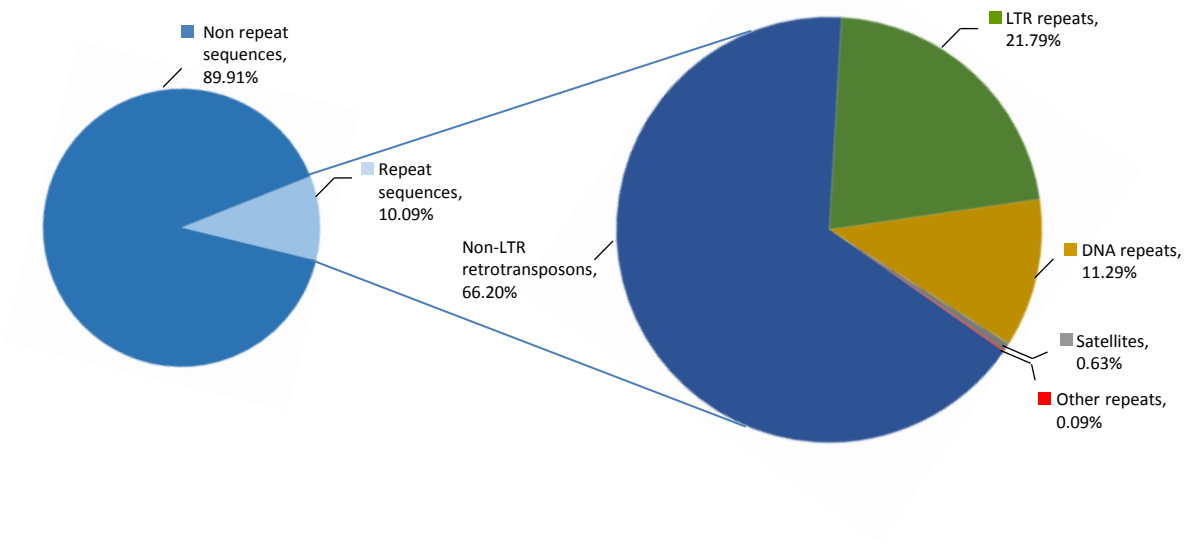


Figure 1-figure supplement 1: 10% of altORFs are present in different classes of repeats.

More than half of the human genome is composed of repeated sequences, and only 10.09% of altORFs are located inside these repeats. These altORFs are detected in non-LTR retrotransposons, LTR repeats, DNA repeats, satellites and other repeats. Proportions were determined using RepeatMasker (version 3.3.0).

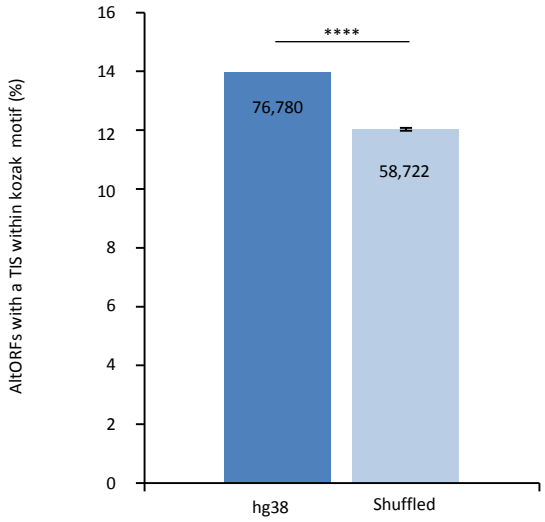


Figure 1-figure supplement 2: The proportion of altORFs with a translation initiation site (TIS) with a Kozak motif in hg38 is significantly different from 100 shuffled hg38 transcriptomes.

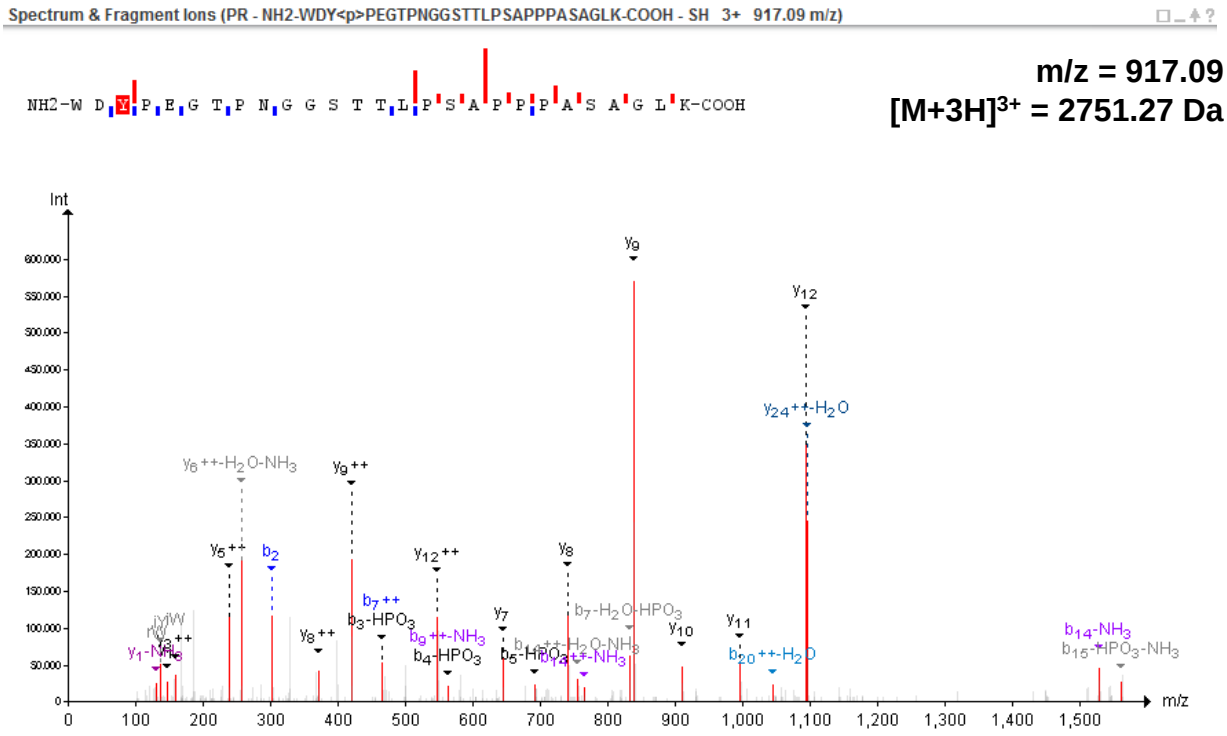
Percentage of altORFs with a TIS within an optimal Kozak sequence in hg38 (dark blue) compared to 100 shuffled hg38 (light blue). Mean and standard deviations for sequence shuffling are displayed, and significant difference was defined by using one sample t test. **** P<0,0001. Note that shuffling all transcripts in the hg38 transcriptome generates a total of 489,073 altORFs on average, compared to 551,380 altORFs in hg38. Most transcripts result from alternative splicing and there are 183,191 unique altORFs in the hg38 transcriptome, while the 489,073 altORFs in shuffled transcriptomes are all unique. Figure 1g shows the percentage of unique altORFs with a kozak motif (15%), while the current Fig. shows the percentage of altORFs with a kozak motif relative to the total number of altORFs (14%).

a

AltLINC01420[nc]

MGDQPCASGRSTLPPGNAREAKPPKKRCLLAPR**WDYPEGTPNGGSTTLPSAPPPASAGLK**SHPPPPEK

b



c

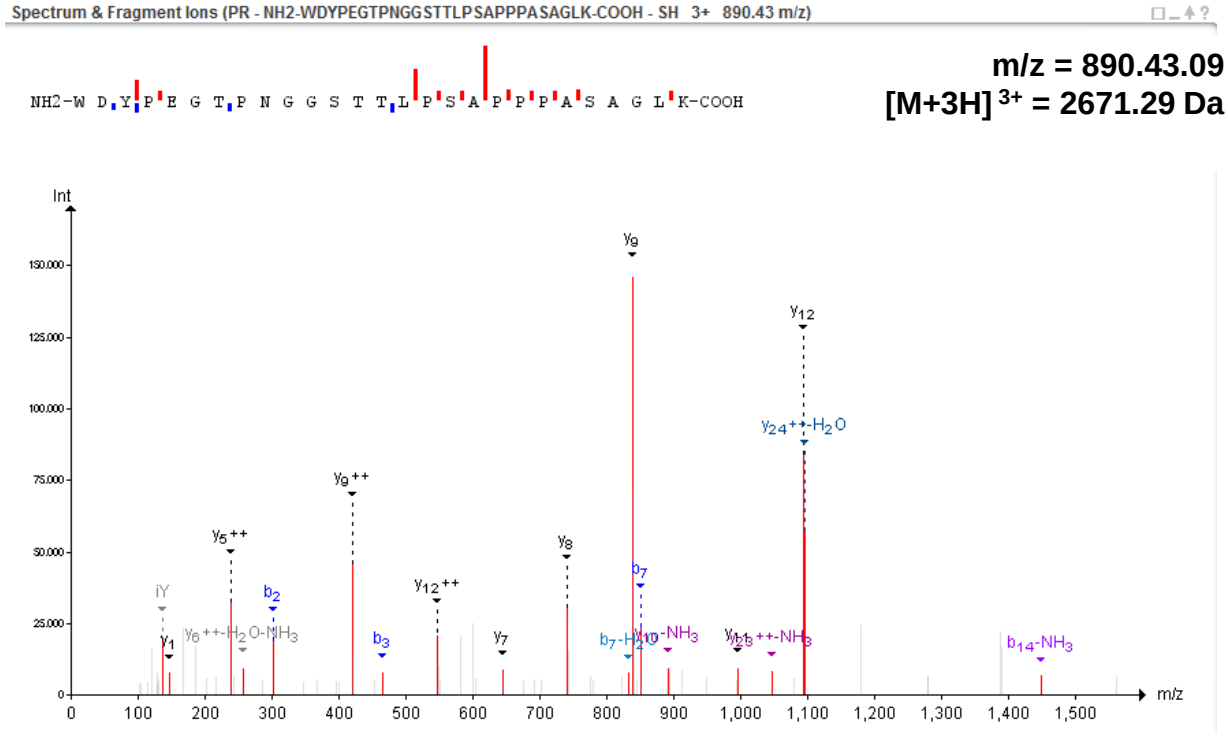


Figure 6-figure supplement 1: Example of a phosphorylated peptide in mitosis - alternative protein AltLINC01420^{nc}.

(**a**) AltLINC01420^{nc} amino acid sequence with detected peptides underlined and phosphorylated peptide in bold (73,9% sequence coverage). (**b**) MS/MS spectrum for the phosphorylated peptide (PeptideShaker graphic interface output). The phosphorylation site is the tyrosine residue, position 2. (**c**) MS/MS spectrum for the non-phosphorylated peptide. The mass difference between the precursor ions between both spectra corresponds to that of a phosphorylation, confirming the specific phosphorylation of this residue in mitosis.

a.

AltTFAM$^{3'}$

MSYINISGQMQHRKHLCSYPAGKFPLSSFNINYPYYFILNIHIIPSQIYWEVQC

b.

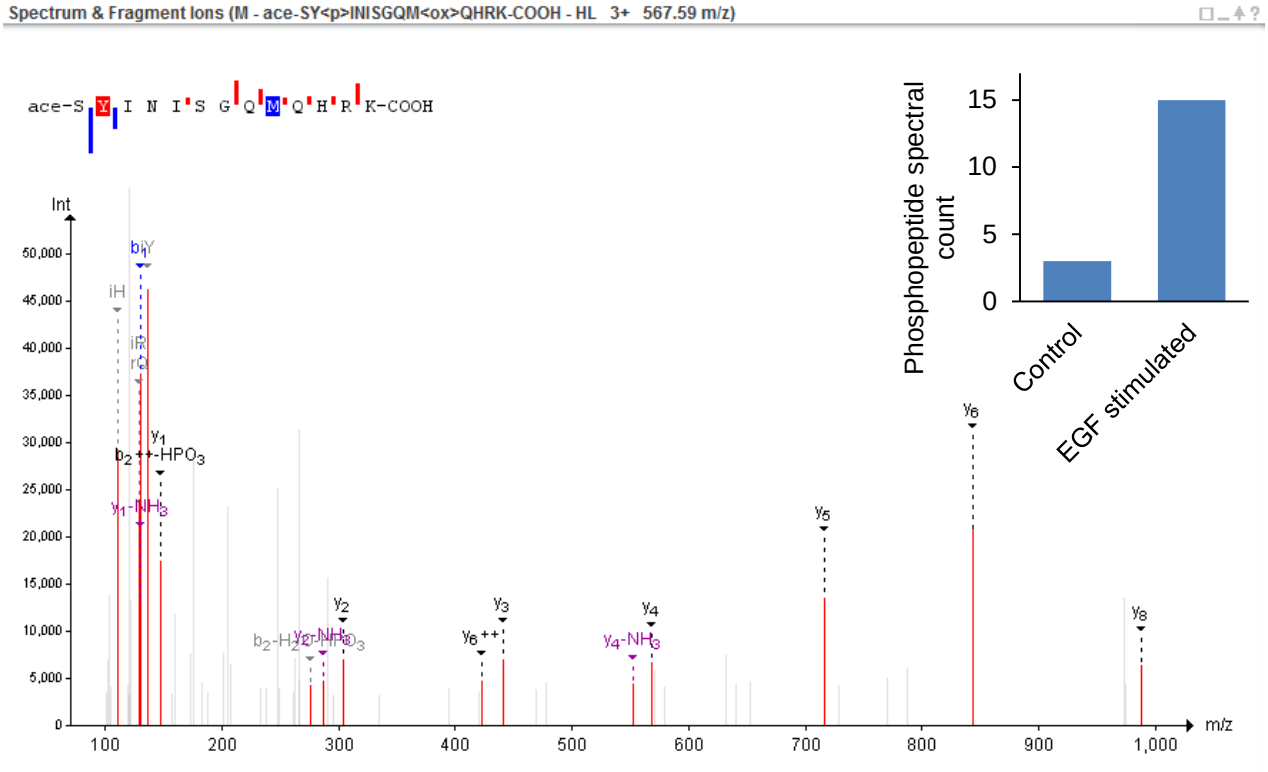


Figure 6-figure supplement 2: Example of a phosphorylated peptide in EGF-treated cells - alternative protein AltTFAM$^{3'}$.

(**a**) AltTFAM$^{3'}$ amino acid sequence with the detected phosphorylated peptide underlined (22,2% sequence coverage). (**b**) MS/MS spectrum for the phosphorylated peptide (PeptideShaker graphic interface output). The phosphorylation site is a tyrosine residue, position 2. The difference in spectral counting indicates an increase in phosphorylation in cells stimulated with EGF.

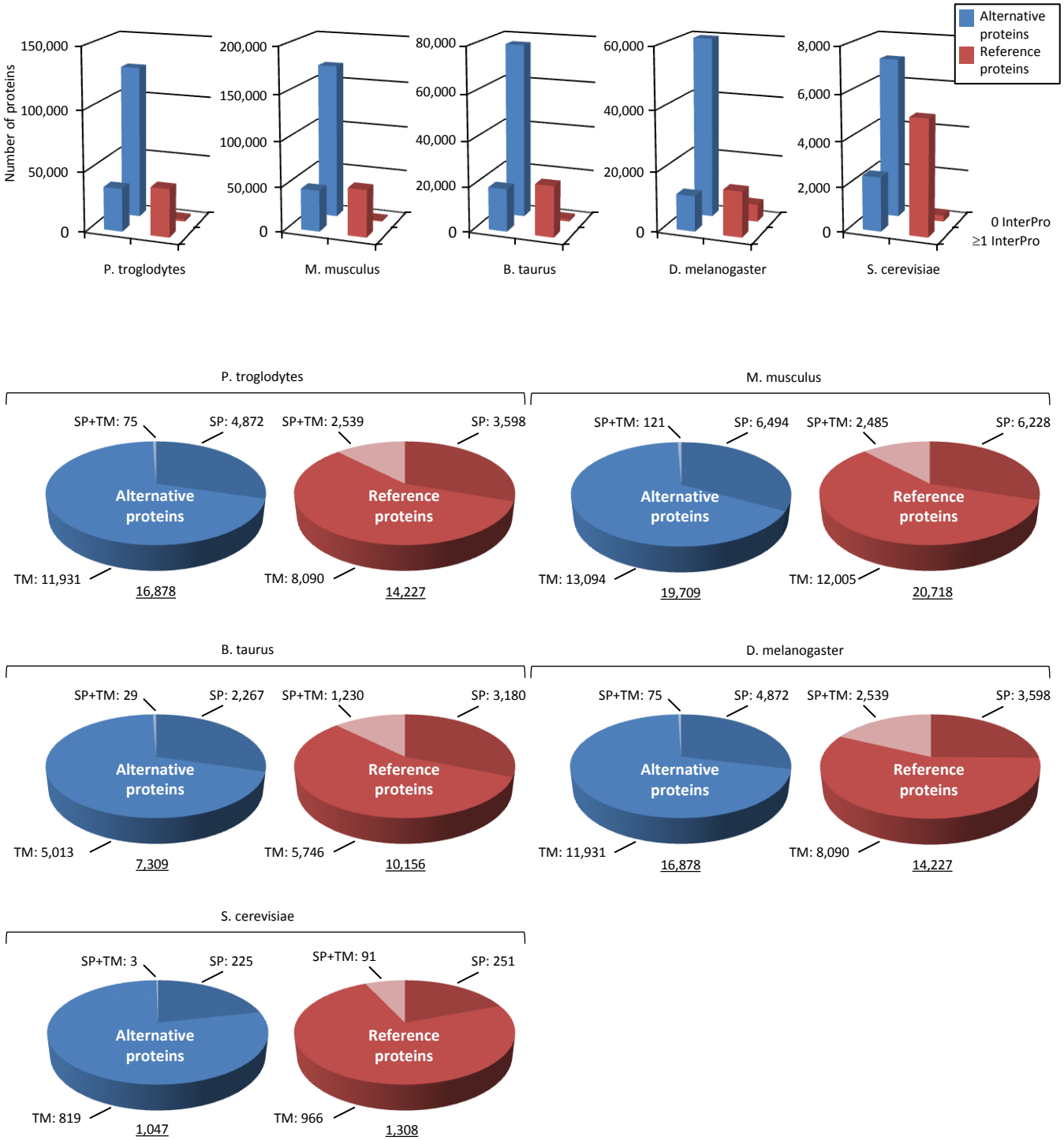


Figure 8-figure supplement 1: Alternative proteome sequence analysis and classification in *P. troglodytes*, *M. musculus*, *B. Taurus*, *D. melanogaster* and *S. cerevisiae*.
For each organism, the number of InterPro signatures (top graphs) and proteins with transmembrane (TM), signal peptide (SP), or TM+SP features (bottom pie charts) is indicated for alternative and reference proteins.

Figure 12-figure supplement 1: Matrix of co-occurrence of InterPro entries between alternative/reference protein pairs coded by the same transcript.
Pixels show the number of times entries co-occur in reference and alternative proteins. Blue pixels indicate that these domains are not shared, white pixels indicate that they are shared once, and red that they are shared twice or more.

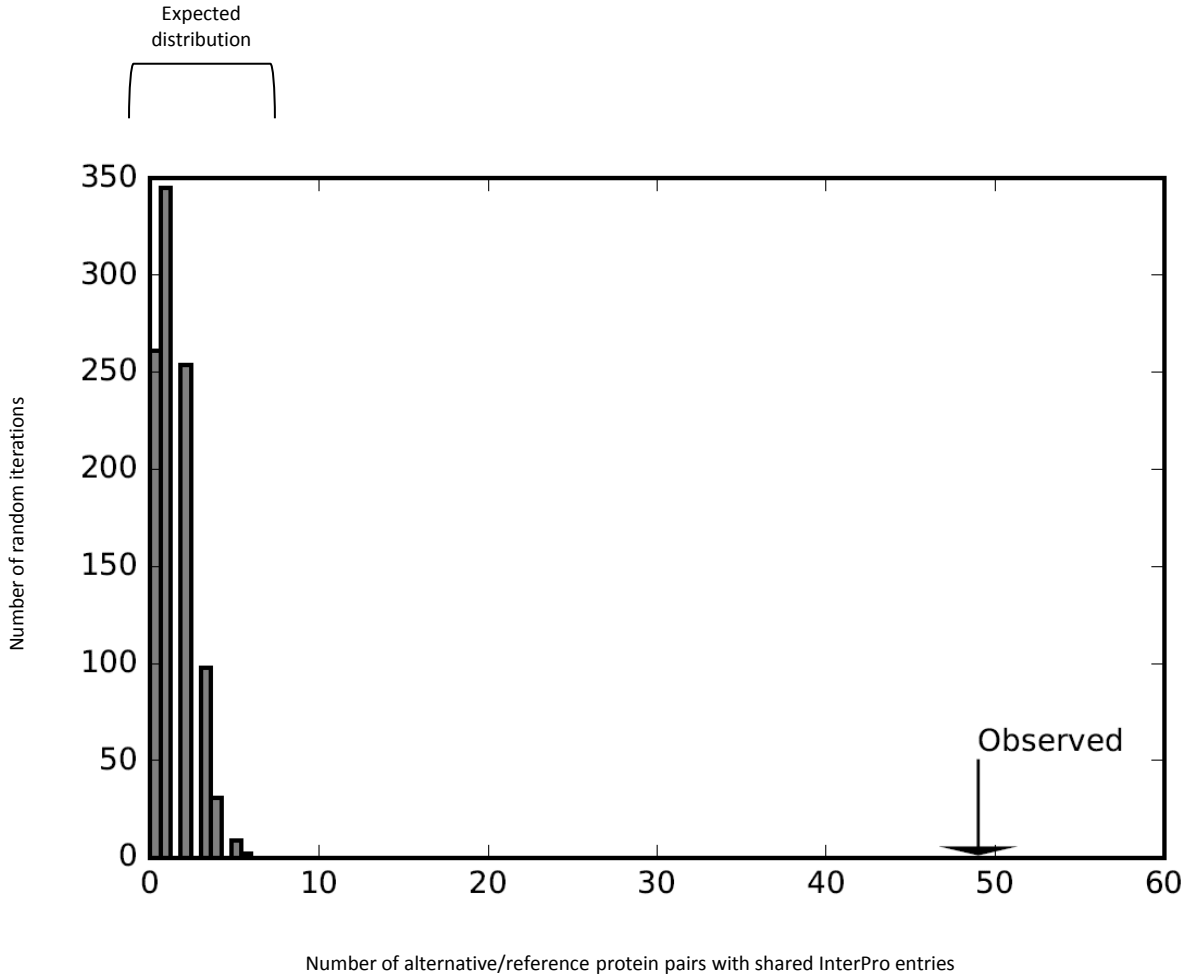
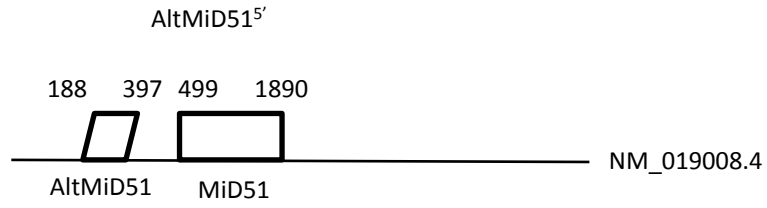
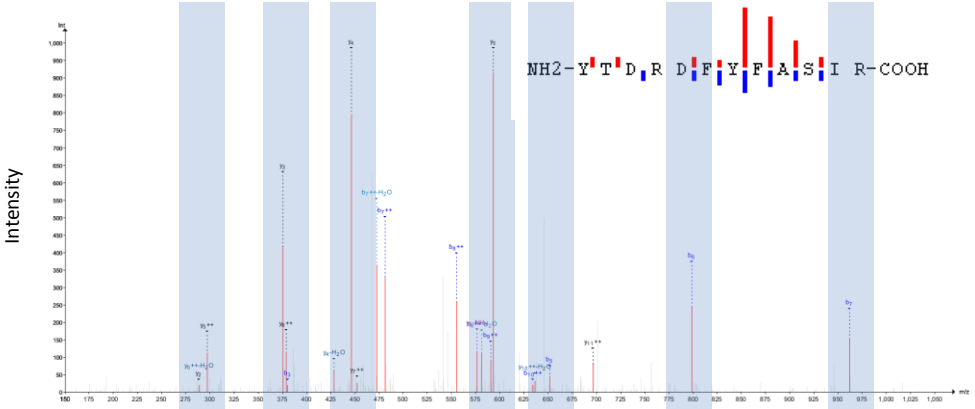


Figure 12-figure supplement 2: Reference and alternative proteins share functional domains.

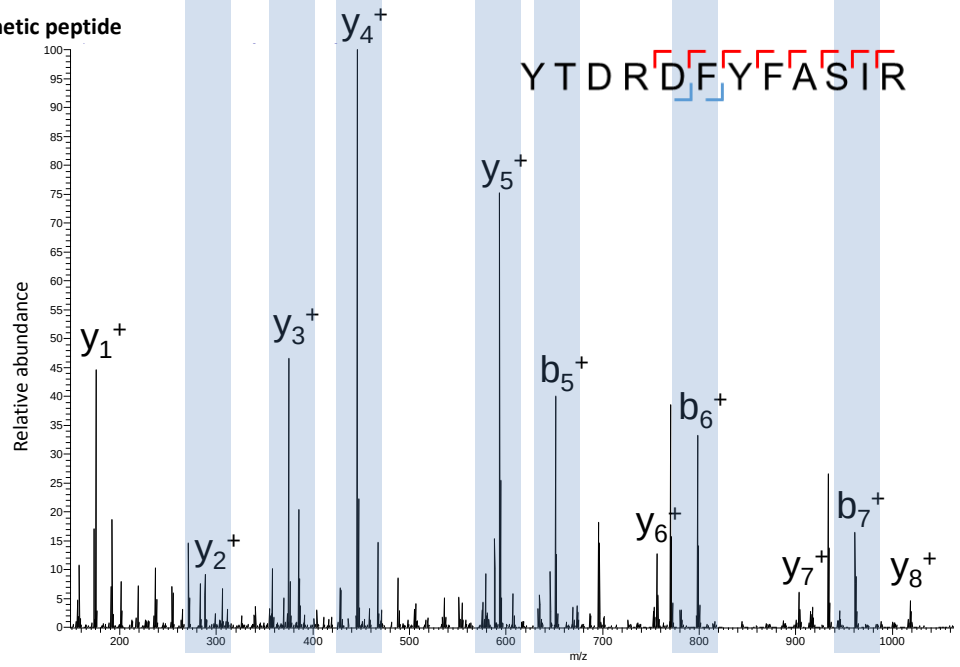
The number of reference/alternative protein pairs that share domains (n = 49) is higher than expected by chance alone (p<0.001). The distribution of expected pairs sharing domains and the observed number are shown. This is the same analysis as the one presented in figure 12b, with the zinc finger domains taken out.



a. Experimental peptide



b. Synthetic peptide



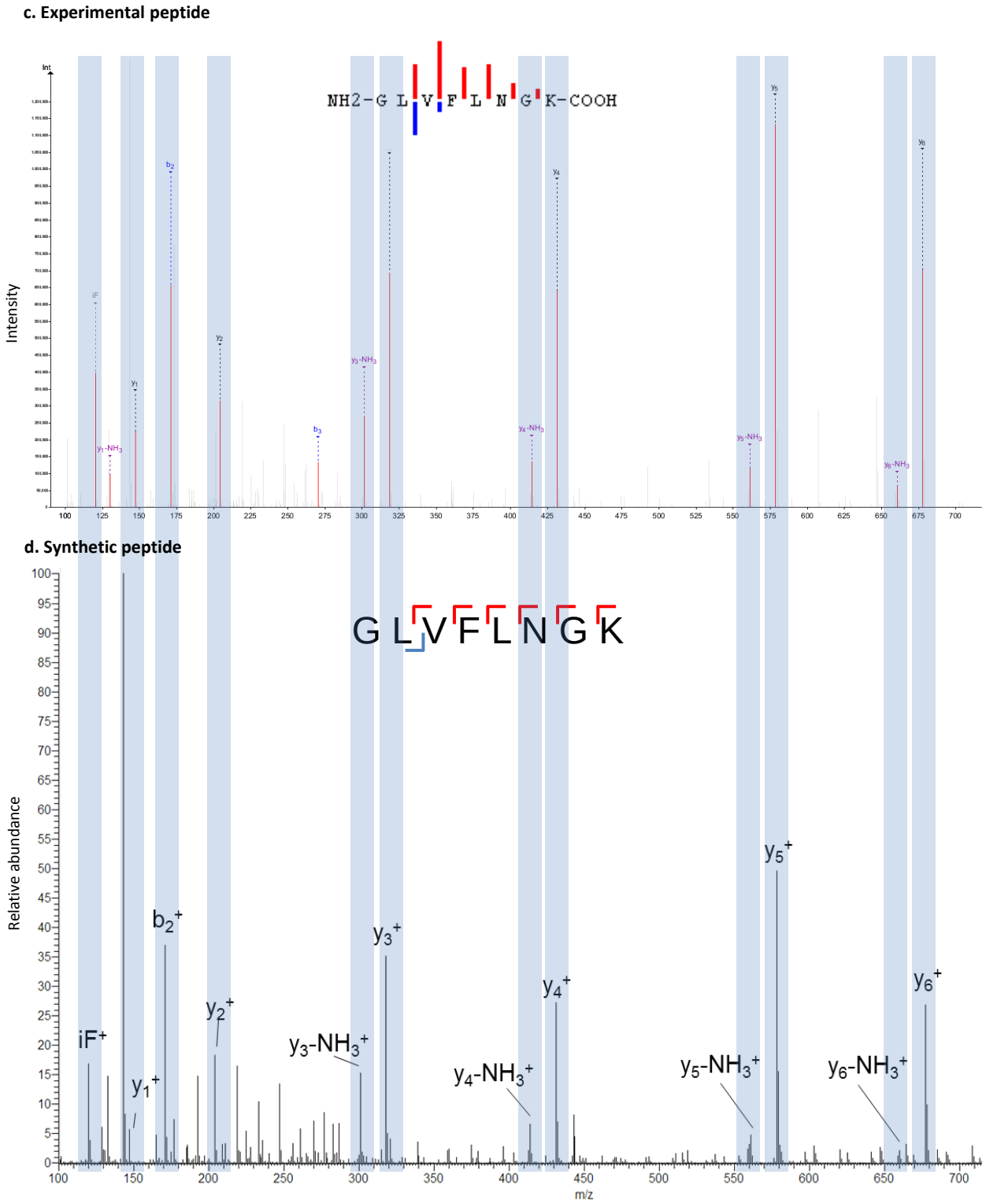


Figure 13-figure supplement 1: Spectra validation for altMiD51.
Example of validation for altMiD51 specific peptides YTDRDFYFASIR and GLVFLNGK. (**a,c**) Experimental MS/MS spectra (PeptideShaker graphic interface output). (**b,d**) MS/MS spectra of the synthetic peptides.
Matching peaks are shown with blue masks. A diagram of the transcript with its accession number and the localization of the altORF and the CDS is shown at the top.

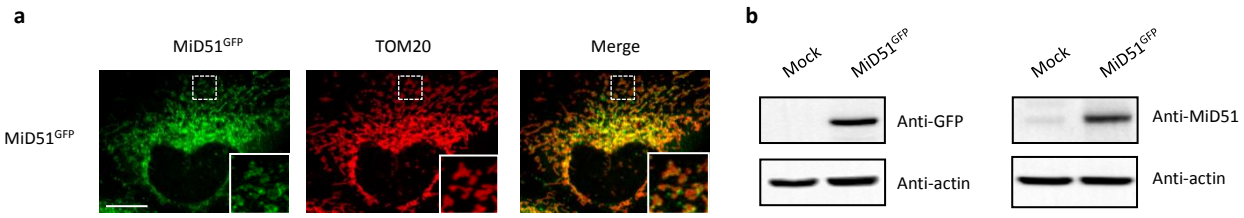


Figure 13-figure supplement 2: MiD51 expression results in mitochondrial fission. (**a**) Confocal microscopy of HeLa cells transfected with MiD51^{GFP} immunostained with anti-TOM20 (red channel) monoclonal antibodies. In each image, boxed areas are shown at higher magnification in the bottom right corner. The localization of MiD51 in fission sites is shown in merged higher magnification inset. Scale bar, 10 μm. (**b**) Human HeLa cells transfected with empty vector (mock) or MiD51^{GFP} were lysed and analyzed by western blot to confirm MiD51^{GFP} expression.

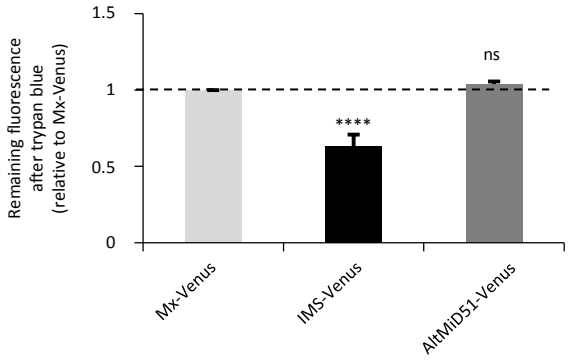


Figure 13-figure supplement 3: AltMiD51 is localized in the mitochondrial matrix. Trypan blue quenching experiment performed on HeLa cells stably expressing the indicated constructs. The fluorescence remaining after quenching by trypan blue is shown relative to Matrix-Venus (Mx-Venus) indicated by the dashed line. (**** p <0,0001, one-way ANOVA). The absence of quenching of the fluorescence compared to IMS-Venus indicates the matricial localization of altMiD51. n≥3 cells were quantified per experiment, and results are from 6 independent experiments. Data are mean ±SEM.

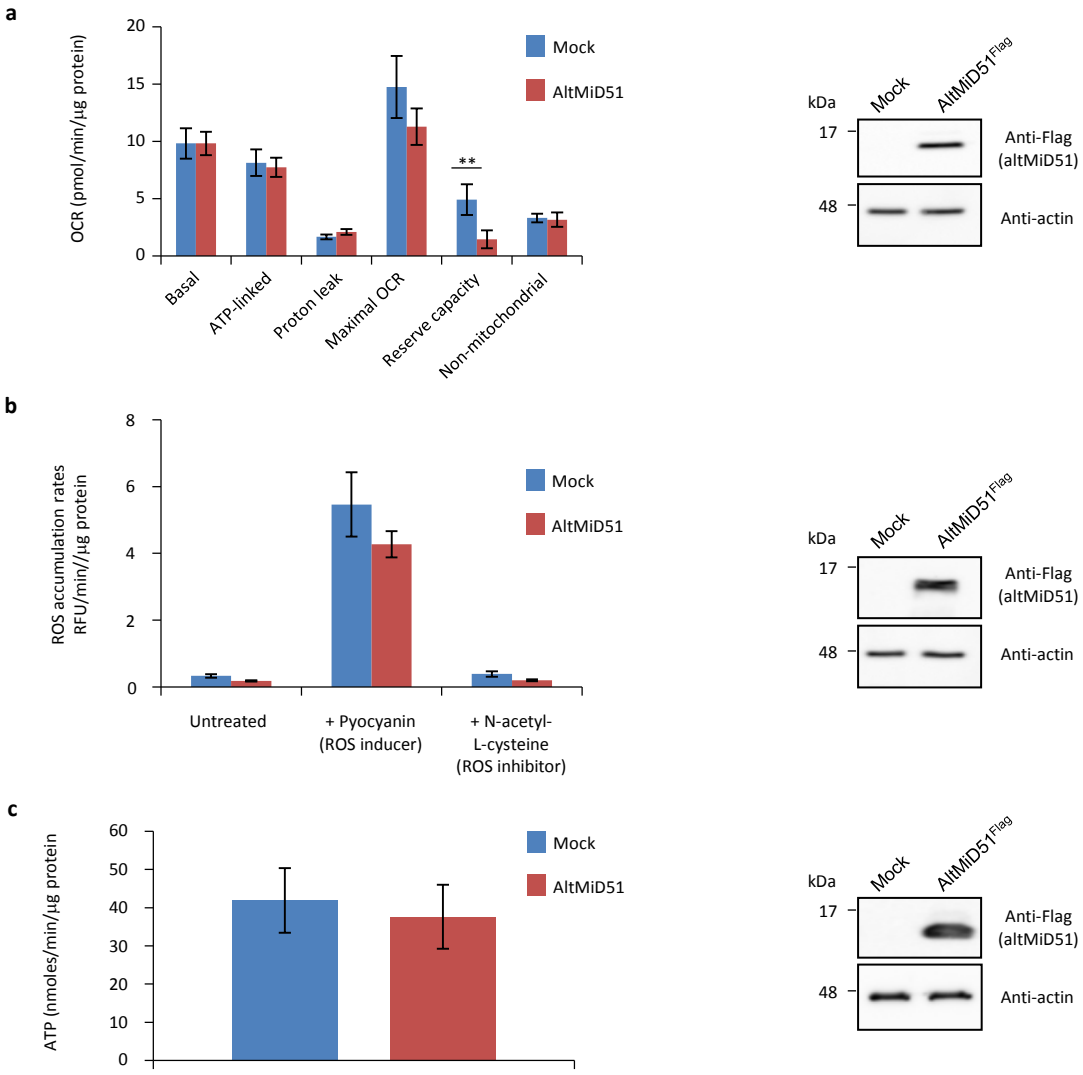


Figure 13-figure supplement 4: Mitochondrial function parameters.

(**a**) Oxygen consumption rates (OCR) in HeLa cells transfected with empty vector (mock) or altMiD51^{Flag}. Mitochondrial function parameters were assessed in basal conditions (basal), in the presence of oligomycin to inhibit the ATP synthase (oxygen consumption that is ATP-linked), FCCP to uncouple the mitochondrial inner membrane and allow for maximum electron flux through the respiratory chain (maximal OCR), and antimycin A/rotenone to inhibit complex III (non-mitochondrial). The balance of the basal OCR comprises oxygen consumption due to proton leak and nonmitochondrial sources. The mitochondrial reserve capacity (maximal OCR- basal OCR) is an indicator of rapid adaptation to stress and metabolic changes. Mean values of replicates are plotted with error bars corresponding to the 95% confidence intervals. Statistical significance was estimated using a two-way ANOVA with Tukey's post-hoc test (**p = 0,004). (**b**) ROS production in mock and altMiD51-expressing cells. Cells were untreated, treated with a ROS inducer or a ROS inhibitor. Results represent the mean value out of three independent experiments, with error bars corresponding to the standard error of the mean (s.e.m.). Statistical significance was estimated using unpaired T-test. (**c**) ATP synthesis rate in mock and altMiD51-expressing cells. No significant differences in ATP production were observed between mock and altMiD51 transfected cells. Results represent the mean of mitochondrial ATP production out of three independent experiments. Error bars represent the standard error of the mean.

At the end of the experiments, cells were collected and proteins analyzed by western blot with antibodies against the Flag tag (altMiD51) or actin, as indicated, to verify the expression of altMiD51. A representative western blot is shown on the right. Molecular weight markers are shown on the left (kDa).

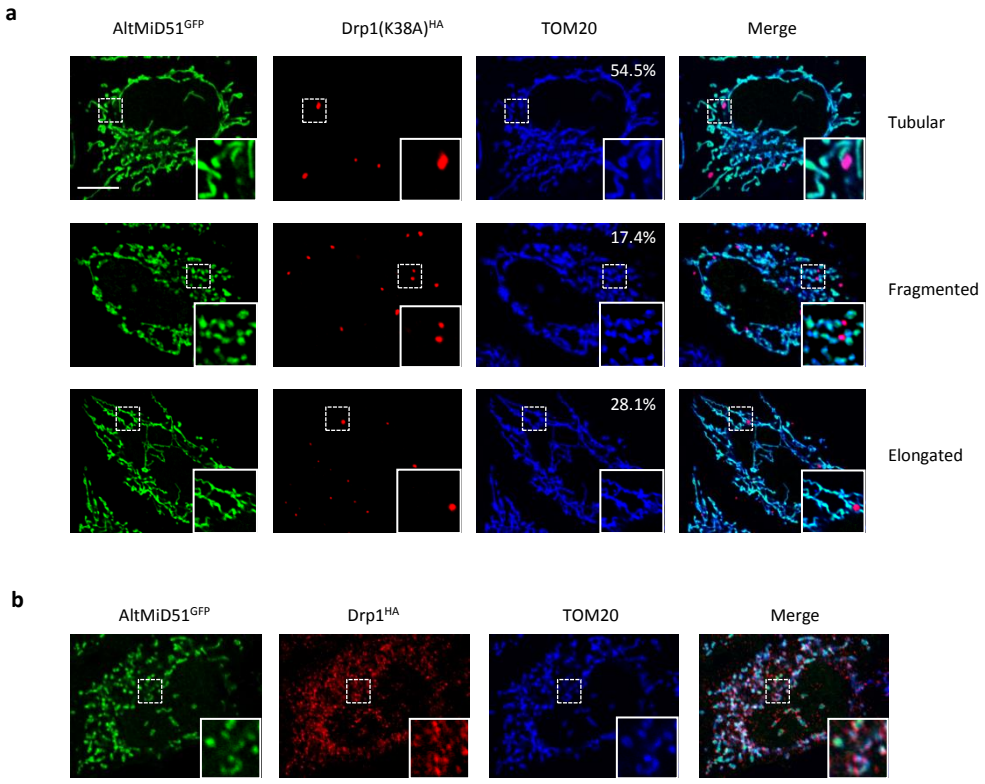


Figure 13-figure supplement 5: Representative confocal images of cells co-expressing altMiD51^{GFP} and Drp1(K38A)^{HA}.

(**a**) Confocal microscopy of HeLa cells co-transfected with altMiD51^{GFP} and Drp1(K38A)^{HA} immunostained with anti-TOM20 (blue channel) and anti-HA (red channel) monoclonal antibodies. In each image, boxed areas are shown at higher magnification in the bottom right corner. % of cells with the indicated morphology is indicated on the TOM20 panels. (**b**) Confocal microscopy of HeLa cells co-transfected with altMiD51^{GFP} and Drp1(wt)^{HA} immunostained with anti-TOM20 (blue channel) and anti-HA (red channel) monoclonal antibodies. In each image, boxed areas are shown at higher magnification in the bottom right corner. Scale bar, 10 μm.

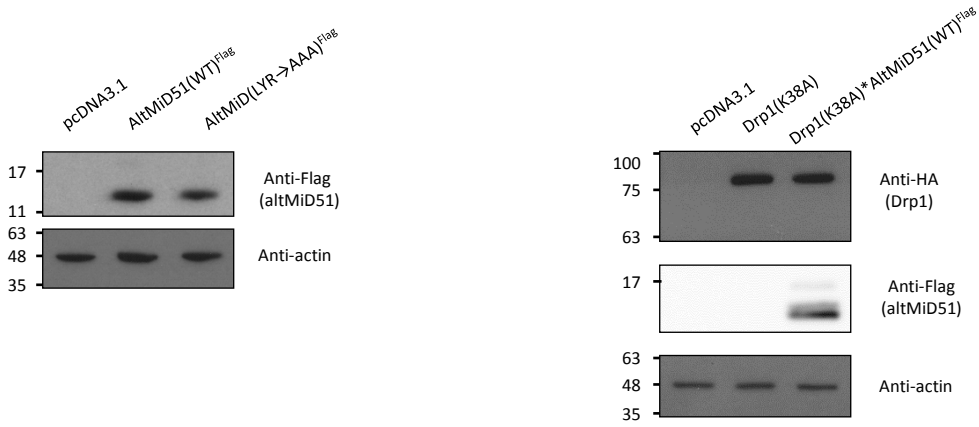


Figure 13-figure supplement 6: Protein immunoblot showing the expression of different constructs in HeLa cells.

HeLa cells were transfected with empty vector (pcDNA3.1), altMiD51(WT)^{Flag}, altMID51(LYR→AAA)^{Flag}, Drp1(K38A)^{HA}, or Drp1(K38A)^{HA} and altMiD51(WT)^{Flag}, as indicated. Proteins were extracted and analyzed by western blot with antibodies against the Flag tag (altMiD51), the HA tag (Drp1K38A) or actin, as indicated. Molecular weight markers are shown on the left (kDa).

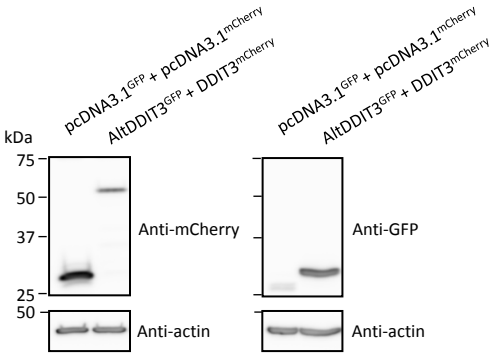
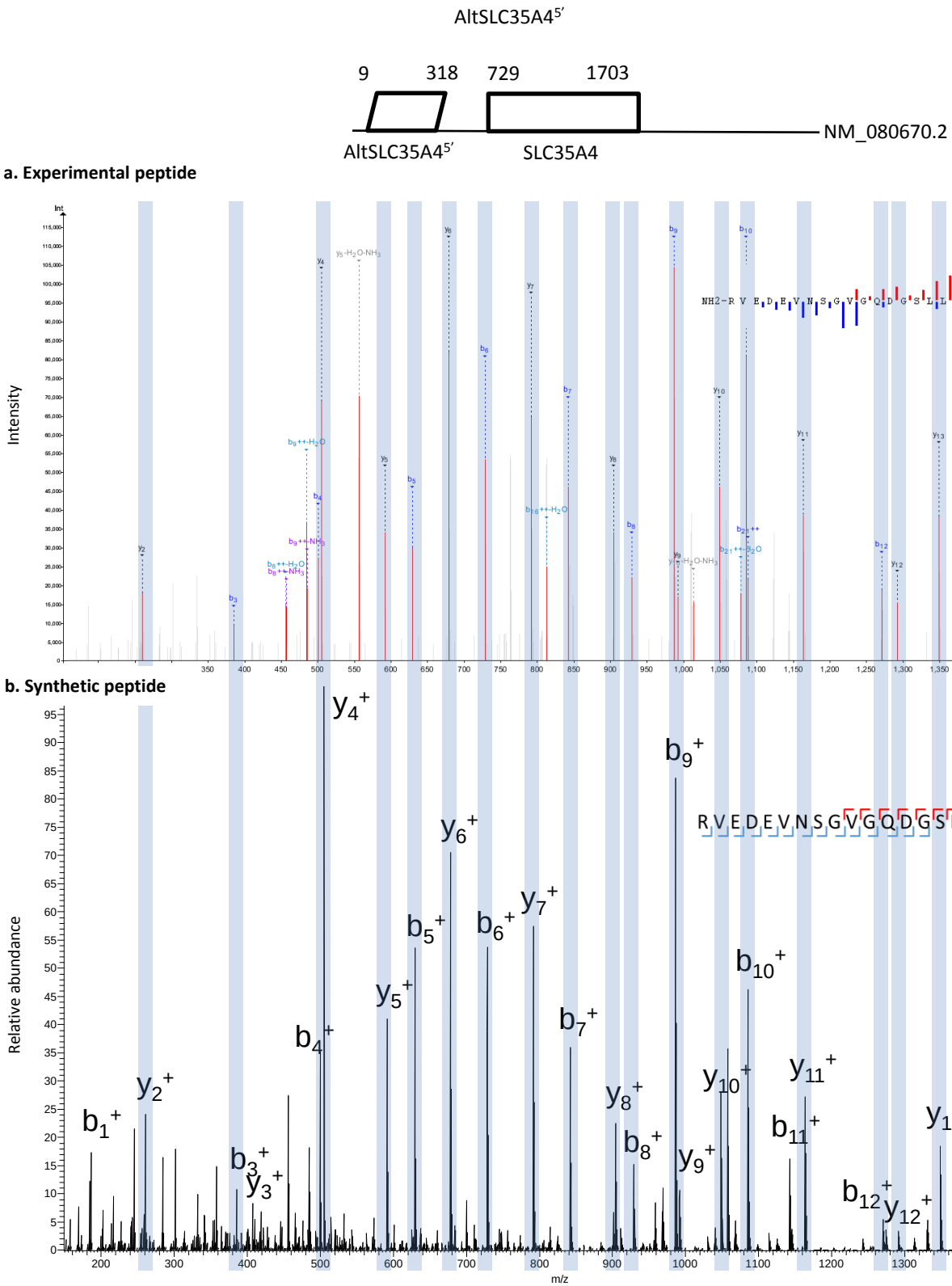


Figure 15-figure supplement 1: Protein immunoblot showing the expression of different constructs in HeLa cells.

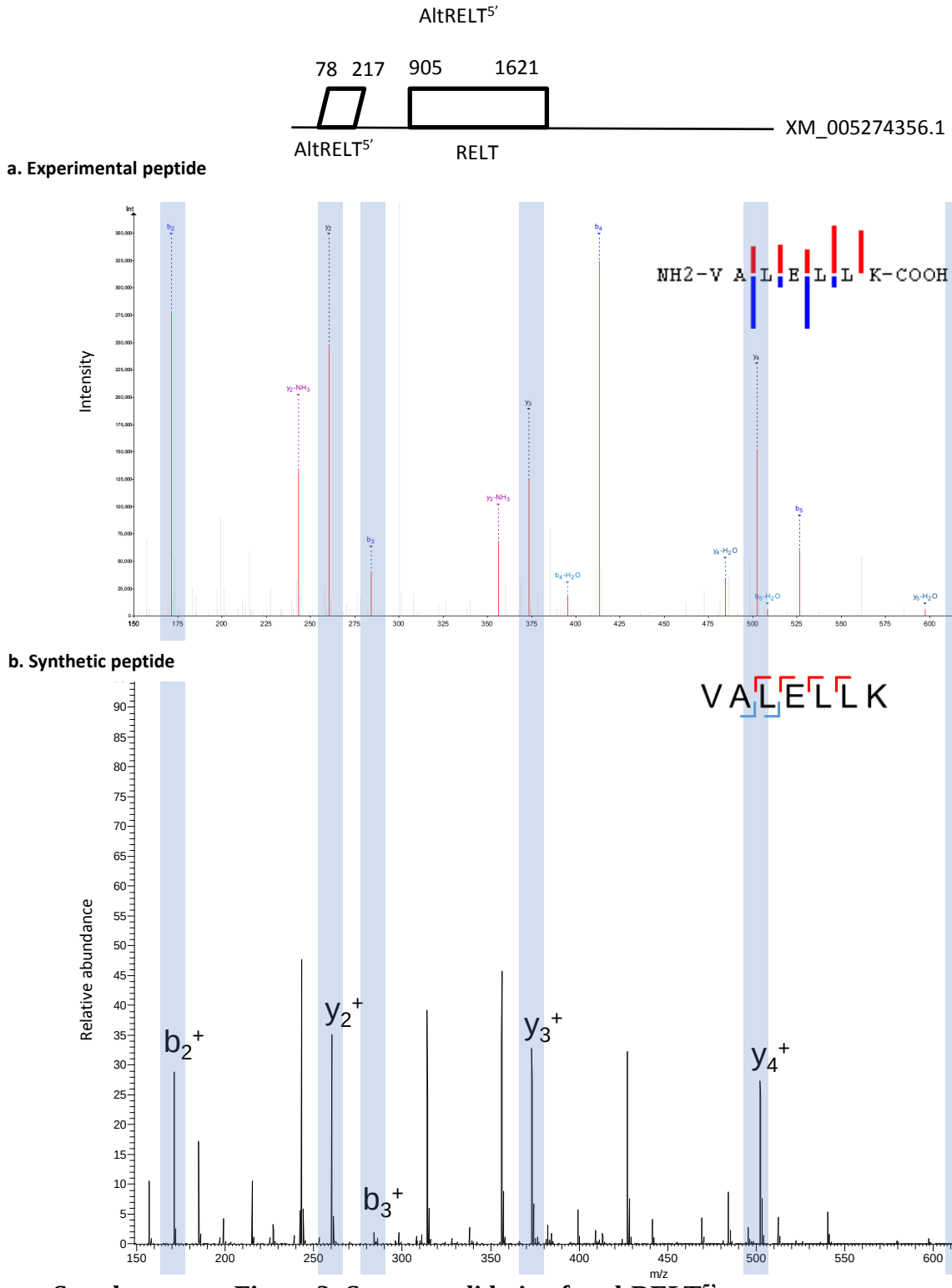
HeLa cells were co-transfected with GFP and mCherry, or altDDIT3^{GFP} and DDIT3^{mCherry}, as indicated. Proteins were extracted and analyzed by western blot with antibodies, as indicated. Molecular weight markers are shown on the left (kDa). AltDDIT3 has a predicted molecular weight of 4.28 kDa and thus migrates at its expected molecular weight when tagged with GFP (~32 kDa).



Supplementary Figure 1: Spectra validation for altSLC35A4$^{5'}$

Example of validation for altSLC35A4$^{5'}$ specific peptide RVEDEVNSGVGQDGSLLSSPFLK. (**a**) Experimental MS/MS spectra (PeptideShaker graphic interface output). (**b**) MS/MS spectra of the synthetic peptide.

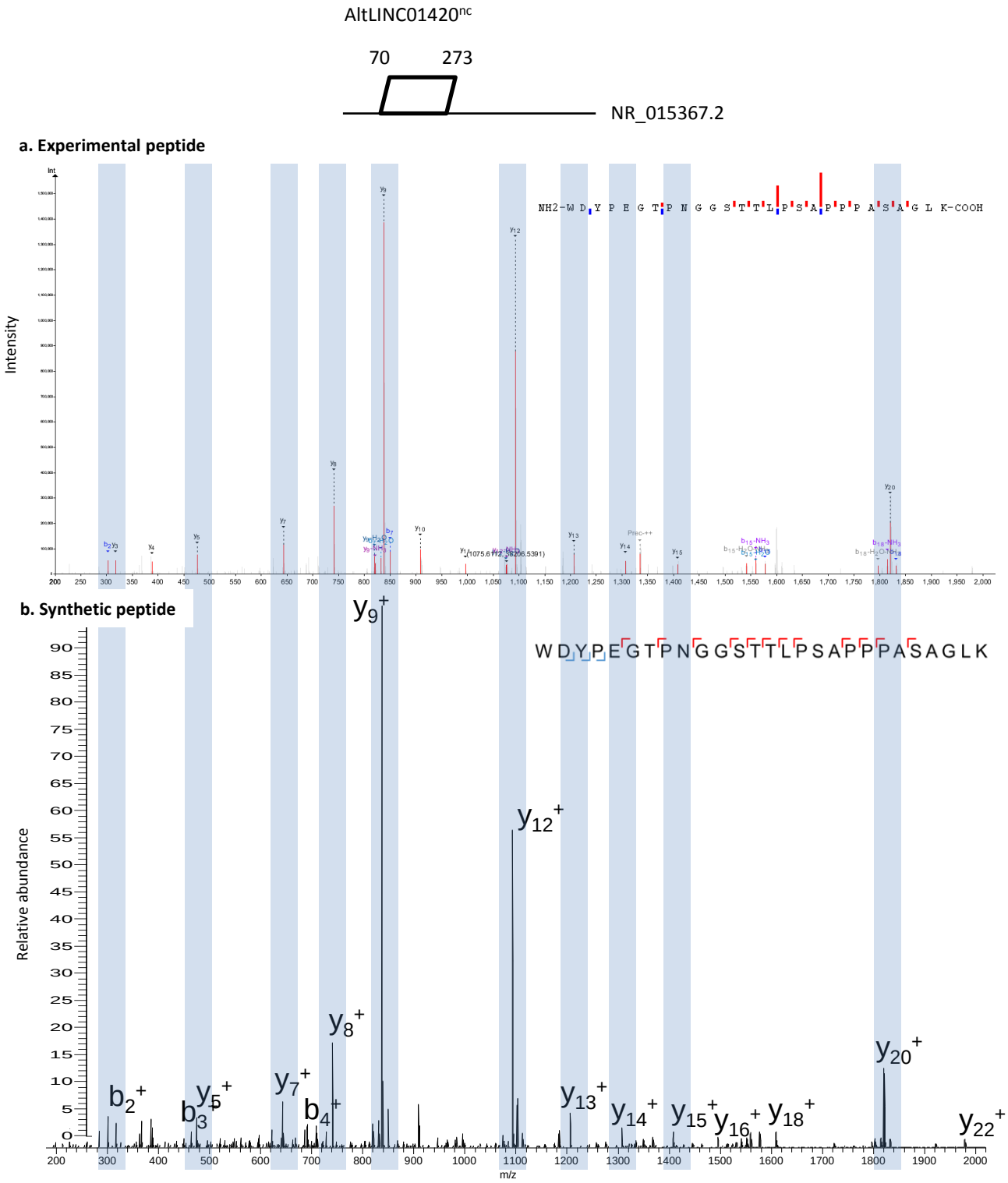
Matching peaks are shown with blue masks. A diagram of the transcript with its accession number and the localization of the altORF and the CDS is shown at the top.



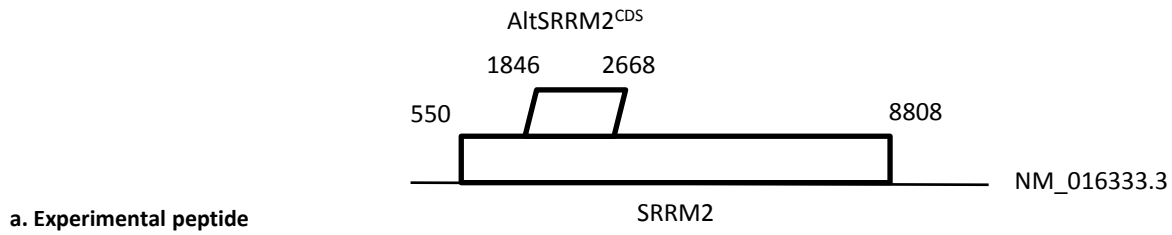
Supplementary Figure 2: Spectra validation for altRELT$^{5'}$

Example of validation for altRELT$^{5'}$ specific peptide VALELLK. (**a**) Experimental MS/MS spectra (PeptideShaker graphic interface output). (**b**) MS/MS spectra of the synthetic peptide.

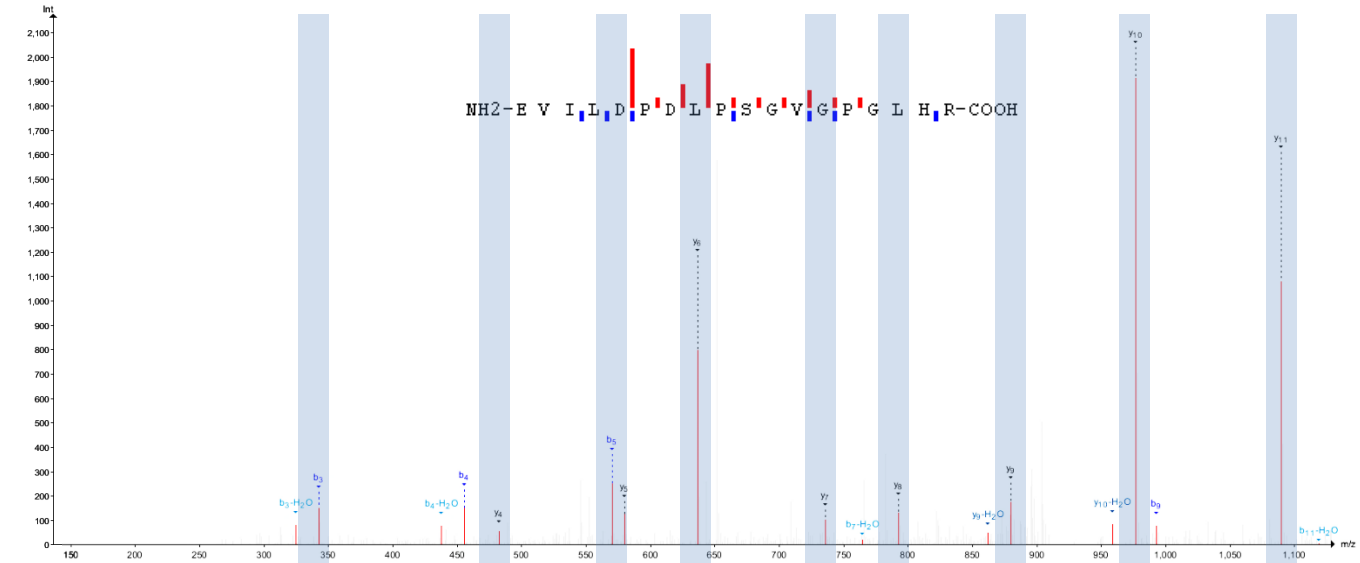
Matching peaks are shown with blue masks. A diagram of the transcript with its accession number and the localization of the altORF and the CDS is shown at the top.



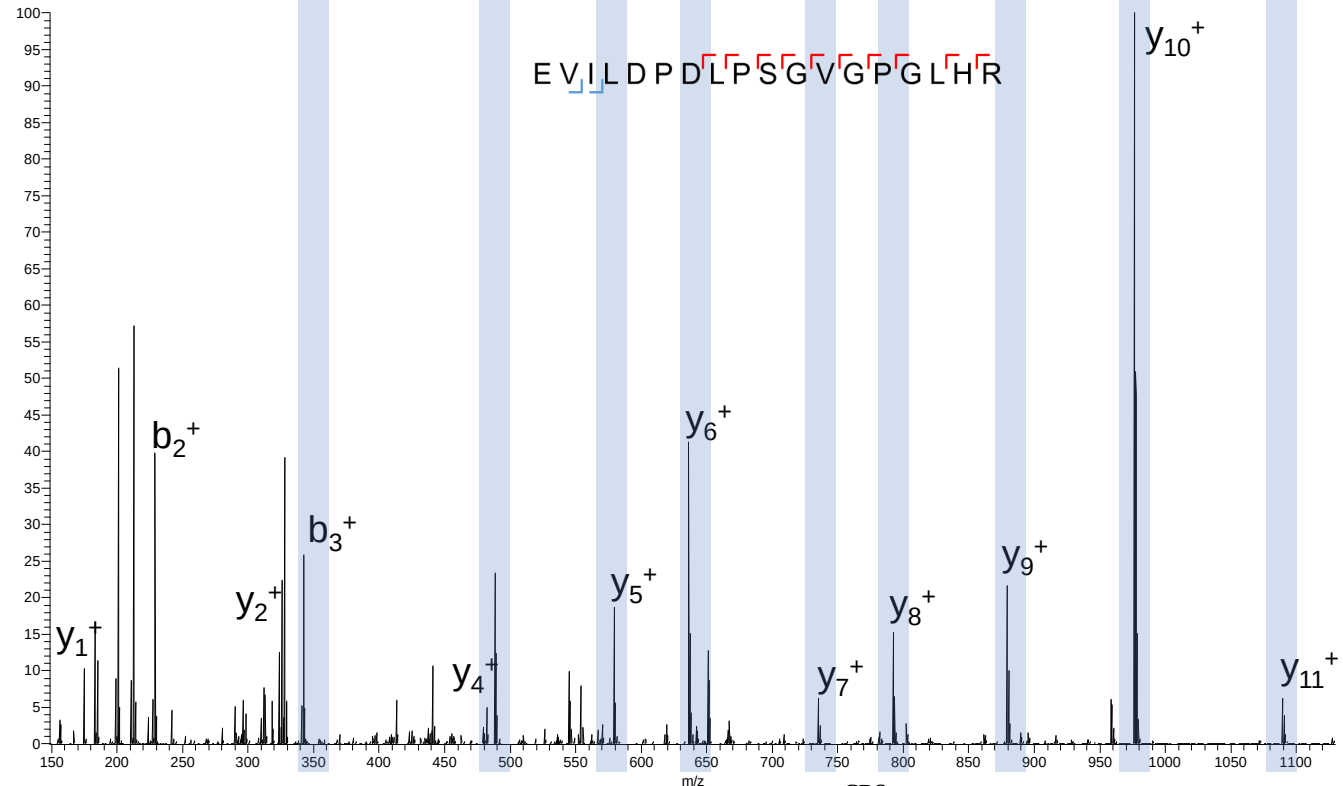
Supplementary Figure 3: Spectra validation for altLINC01420^{nc}
Example of validation for altLINC01420^{nc} specific peptide WDYPEGTPNGGSTTLPSAPPPASAGLK. (**a**) Experimental MS/MS spectra (PeptideShaker graphic interface output). (**b**) MS/MS spectra of the synthetic peptide. Matching peaks are shown with blue masks. A diagram of the transcript with its accession number and the localization of the altORF is shown at the top.



a. Experimental peptide



b. Synthetic peptide



Supplementary Figure 4: Spectra validation for altSRRM2^{CDS}

Example of validation for altSRRM2^{CDS} specific peptide EVILDPDLPSGVGPGLHR. (**a**) Experimental MS/MS spectra (PeptideShaker graphic interface output). (**b**) MS/MS spectra of the synthetic peptide.

Matching peaks are shown with blue masks. A diagram of the transcript with its accession number and the localization of the altORF and the CDS is shown at the top.