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

Sondos Samandi, Annie Roy, Vivian Delcourt, Jean-François Lucier, Jules Gagnon, Maxime Beaudoin, Benoît Vanderperre, Marc-André Breton, Julie Motard, Jean-François Jacques, et al.

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, Québec, Canada; [6]Département de biologie and IBIS, Université Laval, Québec, Canada; [7]PROTEO, Québec Network for Research on Protein Function, Structure, and Engineering, Québec, Canada; [8]Université 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: Département de biochimie (Z8-2001), Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, 3201 Jean Mignault, Sherbrooke, Québec 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 using a database containing predicted alternative proteins. Protein domains and co-conservation analyses suggest a potential functional relationship 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 their roles 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 the F0 subunit 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 new genes, which are thought to contribute to evolutionary innovations. Because these genes generally code for small proteins [19–22], they are difficult to unambiguously detect by proteomics and in fact are 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

important functions and are ultimately one of the most important sources of functional novelty, it is time to address the challenge of their functional annotations[13].

We systematically reanalyzed several eukaryotic transcriptomes to annotate previously unannotated ORFs which we term alternative ORFs (altORFs), and we annotated the corresponding hidden proteome. Here, altORFs are defined as potential protein-coding ORFs in ncRNAs, in UTRs or in different reading frames from annotated CDSs in mRNAs (Figure 1a). For clarity, predicted proteins translated from altORFs are termed alternative proteins and proteins translated from annotated CDSs are termed reference proteins.

Our goal was to provide functional annotations of alternative proteins by (1) analyzing relative patterns of evolutionary conservation between alternative and reference proteins and their corresponding coding sequences; (2) estimating the prevalence of alternative proteins both by bioinformatics analysis and by detection in large experimental datasets; (3) detecting functional signatures in alternative proteins; and (4) predicting and testing functional cooperation between alternative and reference proteins.

Results

Prediction of altORFs and alternative proteins. We predicted a total of 539,134 altORFs compared to 68,264 annotated CDSs in the human transcriptome (Figure 1b, Table 1). Because identical ORFs can be present in different RNA isoforms transcribed from the same genomic locus, the number of unique altORFs and CDSs becomes 183,191

and 54,498, 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, altORFs are on average small, with a size ranging from 30 to 1480 codons. Accordingly, the median size of predicted human 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). 9.83% of altORFs are located in repeat sequences (Figure 1-figure supplement 1a), compared to 2.45% of CDSs. 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 frequencies 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 (A/GNNAUGG), as compared to 19,745 (36% of 54,498) 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 (Figures 2a and c). 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].

Since the same gene may contain a conserved CDS and one or several conserved altORFs, we analyzed the co-conservation of orthologous altORF-CDS pairs. Our results

show a very large fraction of co-conserved alternative-reference protein pairs in several species (Figure 3). Detailed results for all species are presented in Table 2.

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 because of redundancy in the genetic code[27]. In the case of CDS regions overlapping altORFs with a shifted reading frame, the third codon positions of the CDSs are either the first or the second in the altORFs, and should thus also undergo purifying selection. We analyzed conservation of third codon positions of CDSs for 100 vertebrate species for 1,088 altORFs completely nested within and co-conserved across vertebrates (human to zebrafish) with their 889 CDSs from 867 genes (Figure 4). We observed that in regions of the CDS overlapping altORFs, third codon positions were evolving at significantly more extreme speeds (slow or quick) than third codon positions of random control sequences from the entire CDS (Figure 4), reaching up to 67-fold for conservation at $p<0.0001$ and 124-fold for accelerated evolution at $p<0.0001$. We repeated this analysis with the 53,862 altORFs completely nested within the 20,814 CDSs from 14,677 genes, independently of their co-co-conservation. We observed a similar trend, with a 22-fold for conservation at $p<0.0001$, and a 24-fold for accelerated evolution at $p<0.0001$ (Figure 4-figure supplement 1).This is illustrated with three altORFs located within the CDS of *NTNG1*, *RET* and *VTI1A* genes (Figure 5). 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).

157

158 **Evidence of expression of alternative proteins.** We provide two lines of evidence
159 indicating that thousands of altORFs are translated into proteins. First, we re-analyzed
160 detected TISs in publicly available ribosome profiling data[28,29], and found 26,531 TISs
161 mapping to annotated CDSs and 12,616 mapping to altORFs in these studies (Figure 6a;
162 Supplementary file 1). Only a small fraction of TISs detected by ribosomal profiling
163 mapped to altORFs[3'] even if those are more abundant than altORF[5'] relative to shuffled
164 transcriptomes, likely reflecting a recently-resolved technical issue which prevented TIS
165 detection in 3'UTRs [30]. New methods to analyze ribosome profiling data are being
166 developed and will likely uncover more translated altORFs [9]. In agreement with the
167 presence of functional altORFs[3'], cap-independent translational sequences were recently
168 discovered in human 3'UTRs[31]. Second, we re-analyzed proteomic data using our
169 composite database containing alternative proteins in addition to annotated reference
170 proteins (Figure 6b; Supplementary file 2). We selected four studies representing
171 different experimental paradigms and proteomic applications: large-scale [32] and targeted
172 [33] protein/protein interactions, post-translational modifications [34], and a combination of
173 bottom-up, shotgun and interactome proteomics [35] . In the first dataset, we detected 3,957
174 predicted alternative proteins in the interactome of reference proteins[32], providing a
175 framework to uncover the function of these proteins. In a second proteomic dataset
176 containing about 10,000 reference human proteins[35], a total of 549 predicted alternative
177 proteins were detected. Using a phosphoproteomic large data set[34], we detected 384
178 alternative proteins. The biological function of these proteins is supported by the
179 observation that some alternative proteins are specifically phosphorylated in cells

stimulated by the epidermal growth factor, and others are specifically phosphorylated during mitosis (Figure 7; Supplementary file 3). We provide examples of spectra validation (Figure 7-figure supplement 1). A fourth proteomic dataset contained 77 alternative proteins in the epidermal growth factor receptor interactome[33] (Figure 6b). A total of 4,872 different alternative proteins were detected in these proteomic data. The majority of these proteins are coded by altORF^{CDS}, but there are also significant contributions of altORF$^{3'}$, altORF^{nc} and altORF$^{5'}$ (Figure 6c). Overall, by mining the proteomic and ribosomal profiling data, we detected the translation of a total of 17,371 unique alternative proteins. 467 of these alternative proteins were detected by both MS and ribosome profiling (Figure 8), providing a high-confidence collection of 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 9; Figure 9-figure supplement 1). We found 41,511 (23%) human alternative proteins with at least one InterPro signature (Figure 9b). Of these, 37,739 (or 20.6%) are classified as small proteins. Interestingly, the reference proteome has a smaller proportion (840 or 1.68%) 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 (Figures 9c-e; Figure 9-figure supplement 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 (Figures 9c, 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 9e).

We extended the functional annotation using the Gene Ontology. A total of 585 alternative proteins were assigned 419 different InterPro entries, and 343 of them were tentatively assigned 192 gene ontology terms (Figure 10). 15.5% (91/585) 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 (*p*-value = 1.13e-05, Fisher's exact test and chi-square test). Thus, predicted alternative proteins with InterPro entries are more likely to be detected, supporting their functional role. 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 11a). Eighteen of these (17.8%) were detected in public proteomic and ribosome profiling

datasets, a percentage that is similar to reference zinc finger proteins (20.1%) (Figure 6, Table 3). Alternative proteins have between 1 and 23 zinc finger domains (Figure 11b). 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 relationships between reference and alternative proteins coded by the same genes. Since one gene may code 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. The functional associations discussed here are potential functional interactions that do not necessarily imply physical interactions; however, there are a few known examples of functional linkage between different proteins encoded in the same gene (Table 4). If there is a functional relationship, one would expect orthologous alternative- reference protein pairs to be co-conserved more often than expected by chance[40]. Our results show a large fraction of co-conserved alternative-reference protein pairs in several species (Figure 3). Detailed results for all species are presented in Table 2.

Another mechanism that could show functional relationships between alternative and

reference proteins encoded in the same gene 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 11) 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 genes 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 linkage between reference and predicted alternative proteins. Domain sharing remains significant ($p<0.001$) even when the most frequent domains, zinc fingers, are not considered (Figure 12-figure supplement 2).

Recently, the interactome of 118 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 possessing zinc finger domains, thus being functionally connected 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 (Supplementary file 5). Five alternative proteins (IP_168460.1, IP_168527.1, IP_270697.1, IP_273983.1, IP_279784.1) 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 relationships between alternative and reference proteins coded by the same genes, and indicates that there are examples of physical interactions.

Finally, we integrated the co-conservation and expression analyses to produce a high-confidence list of alternative proteins predicted to have a functional relationship with their reference proteins and found 2,715 alternative proteins in mammals (*H. sapiens* to *B. taurus*), and 44 in vertebrates (H. *sapiens* to *D. rerio*) (Supplementary file 6). In order to further test for functional relationship between alternative/reference protein pairs in this list, we focused on alternative proteins detected with at least two peptide spectrum matches or with high TIS reads. From this subset, we selected altMiD51 (IP_294711.1) among the top 2% of alternative proteins detected with the highest number of unique peptides in proteomics studies, and altDDIT3 (IP_211724.1) among the top 2% 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 Kyoto, HeLa S3, THP1 cells and gut tissue), and we validated some spectra using synthetic peptides (Figure 13-figure

supplement 1), and it 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 change oxygen consumption nor ATP and reactive oxygen species production (Figure 13-figure supplement 4). The decrease in spare respiratory capacity in altMiD51-expressing cells (Figure 13-figure supplement 4a) likely resulted from mitochondrial fission[46]. The LYR domain is essential for altMiD51-induced mitochondrial fission since a mutant of the LYR domain, altMiD51(LYR→AAA) was unable to convert the mitochondrial morphology from tubular to fragmented (Figure 13d). Drp1(K38A), a dominant negative mutant of Drp1 [47], largely prevented the ability of altMiD51 to induce mitochondrial fragmentation (Figure 13d; Figure 13-figure supplement 5a). In a control experiment, co-expression of wild-type Drp1 and altMiD51 proteins resulted in mitochondrial fragmentation (Figure 13-figure supplement 5b). Expression of the different constructs

used in these experiments was verified by western blot (Figure 13-figure supplement 6). Drp1 knockdown interfered with altMiD51-induced mitochondrial fragmentation (Figure 14), confirming the proposition that Drp1 mediates altMiD51-induced mitochondrial fragmentation. It remains possible that altMiD51 promotes mitochondrial fission independently of Drp1 and is able to reverse the hyperfusion induced by Drp1 inactivation. However, Drp1 is the key player mediating mitochondrial fission and most likely mediates altMiD51-induced mitochondrial fragmentation, as indicated by our results.

AltDDIT3 is a 31 amino acid alternative protein conserved in vertebrates and co-conserved with its reference protein DDIT3 from human to bovine (Supplementary file 6). Its coding sequence overlaps the end of exon 1 and the beginning of exon 2 of the *DDIT3/CHOP/GADD153* gene. These exons form part of the 5'UTR for the canonical 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}. Expression of these constructs was verified by western blot (Figure 15-figure supplement 1). 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 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; Figure 15-figure supplement 2) [50]. Finally, in lysates from cells co-expressing altDDIT3^{GFP} and DDIT3^{mCherry}, DDIT3^{mCherry} was

341 immunoprecipitated with GFP-trap agarose, confirming an interaction between the small
342 altDDTI3 and the large DDIT3 proteins encoded in the same gene (Figure 15e).
343
344
345 **Discussion**
346 We have provided the first functional annotation of altORFs with a minimum size of 30
347 codons in different genomes. The comprehensive annotation of *H sapiens* altORFs is
348 freely available to download at https://www.roucoulab.com/p/downloads (Homo sapiens
349 functional annotation of alternative proteins based on RefSeq GRCh38 (hg38)
350 predictions). In light of the increasing evidence from approaches such as ribosome
351 profiling and MS-based proteomics that the one mRNA-one canonical CDS assumption
352 is strongly challenged, our findings provide the first clear functional insight into a new
353 layer of regulation in genome function. While many observed altORFs may be
354 evolutionary accidents with no functional role, several independent lines of evidence
355 support translation and a functional role for thousands of alternative proteins: (1)
356 overrepresentation of altORFs relative to shuffled sequences; (2) overrepresentation of
357 altORF Kozak sequences; (3) active altORF translation detected via ribosomal profiling;
358 (4) detection of thousands of alternative proteins in multiple existing proteomic datasets;
359 (5) correlated altORF-CDS conservation, but with overrepresentation of highly conserved
360 and fast-evolving altORFs; (6) overrepresentation of identical InterPro signatures
361 between alternative and reference proteins encoded in the same mRNAs; (7) several
362 thousand co-conserved paired alternative-reference proteins encoded in the same gene;
363 and (8) presence of clear, striking examples in altMiD51, altDDIT3 and 5 alternative

proteins interacting with their reference zinc finger proteins. While 5 of these 8 lines of evidence support an unspecified functional altORF role, 4 of them (5, 6, 7 and 8) 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.

The presence of different coding sequences in the same gene provides a coordinated transcriptional regulation. Consequently, the transcription of different coding sequences can be turned on or off together, similar to prokaryotic operons. Thus, the observation that alternative-reference protein pairs encoded in the same genes have functional relationships is not completely unexpected. This observation is also in agreement with increasing evidence that small proteins regulate the function of larger proteins[51]. We speculate that clustering of a CDS and one or more altORFs in the same genes, and thus in the same transcription unit allows cells to adapt more quickly with optimized energy expenditure to environmental changes.

Upstream ORFs here labeled altORFs$^{5'}$ are important translational regulators of canonical CDSs in vertebrates[52]. Interestingly, the altORF$^{5'}$ encoding altDDIT3 was characterized as an inhibitory upstream ORF [53,54], but evidence of endogenous expression of the corresponding small protein was not sought. The detection of altMiD51 and altDDIT3 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 [55,56]. 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 585 alternative 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*[20,57,58] 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, evolutionarily novel genes tend to be poorly expressed, again reducing the probability of detection [20]. Here, we used a combination of five search engines, thus increasing the confidence and sensitivity of hits compared to single-search-engine processing[59,60]. 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.

Only a relatively small percentage of alternative proteins (22.6%) are functionally annotated with Interpro signatures, compared to reference proteins (96.9%). An obvious explanation is the small size of alternative proteins with a median size of 45 amino acids, which may not be able to accommodate large domains. It has been proposed that small proteins may be precursors of new proteins but require an elongation of their coding sequence before they display a useful cellular activity [19,51]. According to this hypothesis, it is possible that protein domains appear only after elongation of the coding sequence. Alternatively, InterPro domains were identified by investigating the reference proteome, and alternative proteins may have new domains and motifs that remain to be characterized. Finally, an unknown fraction of predicted altORFs may not be translated or may code for non-functional peptides.

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[61–65]. To determine whether or not

433 this functional cooperation is a general feature of small/large protein pairs encoded in the
434 same gene will require much more experimental evidence, but our results strongly
435 support this hypothesis.
436

437 ## Materials and methods

438 **Generation of alternative open reading frames (altORFs) and alternative protein**
439 **databases.** Throughout this manuscript, annotated protein coding sequences and proteins
440 in current databases are labelled annotated coding sequences or CDSs and reference
441 proteins, respectively. For simplicity reasons, predicted alternative protein coding
442 sequences are labelled alternative open reading frames or altORFs.
443 To generate MySQL databases containing the sequences of all predicted alternative
444 proteins translated from reference annotation of different organisms, a computational
445 pipeline of Perl scripts was developed as previously described with some modifications[37].
446 Genome annotations for *H. sapiens* (release hg38, Assembly: GCF_000001405.26), *P.*
447 *troglodytes* (Pan_troglodytes-2.1.4, Assembly: GCF_000001515.6), *M. musculus*
448 (GRCm38.p2, Assembly: GCF_000001635.22), *D. melanogaster* (release 6, Assembly:
449 GCA_000705575.1), *C. elegans* (WBcel235, Assembly: GCF_000002985.6) and *S.*
450 *cerevisiae* (Sc_YJM993_v1, Assembly: GCA_000662435.1) were downloaded from the
451 NCBI website (http://www.ncbi.nlm.nih.gov/genome). For *B. taurus* (release UMD
452 3.1.86), *X. tropicalis* (release JGI_4.2) and *D. rerio* (GRCz10.84), genome annotations
453 were downloaded from Ensembl (http://www.ensembl.org/info/data/ftp/). Each annotated
454 transcript was translated *in silico* with Transeq[66]. All ORFs starting with an AUG and
455 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 when overlapping the 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[67].

Identification of TISs. The global aggregates of initiating ribosome profiles data were obtained from the initiating ribosome tracks in the GWIPS-viz genome browser[28] with ribosome profiling data collected from five large scale studies[2,9,68–70]. 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 (RRID:SCR_001173). 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

479 altORF TIS). An additional recent study was also included in our analysis[29]. In this study,
480 a threshold of 5 reads was used. Raw sequencing data for ribosome protected fragments
481 in harringtonine treated cells was aligned to the human genome (GRCh38) using bowtie2
482 (2.2.8)[71]. Similar to the method used in this work, altORFs with at least 5 reads
483 overlapping one position in the kozak region were considered as having an
484 experimentally validated TIS.
485
486 **Generation of shuffled transcriptomes.** Each annotated transcript was shuffled using
487 the Fisher-Yates shuffle algorithm. In CDS regions, all codons were shuffled except the
488 initiation and stop codons. For mRNAs, we shuffled the 5'UTRs, CDSs and 3'UTRs
489 independently to control for base composition. Non-coding regions were shuffled at the
490 nucleotide level. The resulting shuffled transcriptome has the following features
491 compared to hg38: same number of transcripts, same transcripts lengths, same nucleotide
492 composition, and same amino-acid composition for the proteins translated from the
493 CDSs. Shuffling was repeated 100 times and the results are presented with average values
494 and standard deviations. The total number of altORFs is 539,134 for hg38, and an
495 average of 489,073 for shuffled hg38. AltORFs and kozak motifs in the 100 shuffled
496 transcriptomes were detected as described above for hg38.
497
498 **Identification of paralogs/orthologs in alternative proteomes.** Both alternative and
499 reference proteomes were investigated. Pairwise ortholog and paralog relationships
500 between the human proteomes and the proteomes from other species, were calculated
501 using an InParanoid-like approach[72], as described below (RRID:SCR_006801). The

following BLAST (RRID:SCR_001010) 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 (*Homo sapiens* 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[73]. 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[72].

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

525 relatively rare; we thus used the probability that the orthologous alternative protein is
526 paired with any orthologous CDS for our null model. Because this probability is by
527 definition higher than the probability that the altORF occurs on the paired CDS, it is a
528 conservative estimate of co-conservation. We took two approaches to estimating this
529 percentage, and then used whichever was higher for each species pair, yielding an even
530 more conservative estimate. First, we assessed the percentage of orthologous reference
531 proteins under the null supposition that each orthologous alternative protein had an equal
532 probability of being paired with any reference protein, orthologous or not. Second, we
533 assessed the percentage of non-orthologous alternative proteins that were paired with
534 orthologous reference proteins. This would account for factors such as longer CDSs
535 having a higher probability of being orthologous and having a larger number of paired
536 altORFs. For example, between *Homo sapiens* and *Mus musculus*, we found that 22,304
537 of 54,498 reference proteins (40.9%) were orthologs. Of the 157,261 non-orthologous
538 alternative proteins, 106,987 (68%) were paired with an orthologous reference protein.
539 Because 68% is greater than 40.9%, we used 68% as the probability for use in our null
540 model. Subsequently, our model strongly indicates co-conservation (Fig. 3 and Table 2;
541 $p<10^{-6}$ based on 1 million binomial simulations; highest observed random percentage
542 =69%, much lower than the observed 96% co-conservation).
543
544 **Analysis of third codon position (wobble) conservation.** Basewise conservation scores
545 for the alignment of 100 vertebrate genomes including *H. sapiens* were obtained from
546 UCSC genome browser
547 (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP100way/)

(RRID:SCR_012479). 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 (see code file). 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 (RRID:SCR_012954), was used to investigate the extent of altORFs derived from transposable elements[74]. Version 3-3-0 was run with default settings.

Alternative protein analysis using InterProScan (RRID:SCR_005829)

InterProScan combines 15 different databases, most of which use Hidden Markov models for signature identification[75]. 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[76] (RRID:SCR_005737).

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 the number of entries between the alternative and reference proteins as 100*intersection/union. All reference proteins and the corresponding alternative proteins were combined together in each comparison so that all domains of all isoforms for a given reference protein were considered in each comparison. The random distribution of

the number of alternative/reference protein pairs that share at least one InterPro entry was computed by shuffling the alternative/reference protein pairs and calculating how many share at least one InterPro entry. This procedure was repeated 1,000 times. Finally, we compared the number and identity of shared InterPro entries in a two dimensional matrix to illustrate which Interpro entries are shared. In many instances, including for zinc-finger coding genes, InterPro entries in alternative/reference protein pairs tend to be related when they are not identical.

Mass Spectrometry identification. Wrapper Perl scripts were developed for the use of SearchGUI v2.0.11[77] (RRID:SCR_012054) and PeptideShaker v1.1.0[59] (RRID:SCR_002520) on the *Université de Sherbrooke's* 39,168 core high-performance *Mammouth Parallèle 2* computing cluster (http://www.calculquebec.ca/en/resources/compute-servers/mammouth-parallele-ii). SearchGUI was configured to run the following proteomics identification search engines: X!Tandem[78], MS-GF+[79], MyriMatch[80], Comet[81], and OMSSA[82]. SearchGUI parameters were set as follow: maximum precursor charge, 5; maximum number of PTM per peptide, 5; X!Tandem minimal fragment m/z, 140; removal of initiator methionine for Comet, 1. A full list of parameters used for SearchGUI and PeptideShaker is available in Supplementary file 2, sheet 1. For PXD000953 dataset[35], precursor and fragment tolerance were set 0.006 Da and 0.1 Da respectively, with carbamidomethylation of C as a fixed modification and Nter-Acetylation and methionine oxidation as variable modifications. For PXD000788[33] and PXD000612[34] datasets, precursor and fragment tolerance were set to 4.5 ppm and 0.1 Da respectively with carbamidomethylation of

617 cysteine as a fixed modification and Nter-Acetylation, methionine oxidation and
618 phosphorylation of serine, threonine and tyrosine as variable modifications. For
619 PXD002815 dataset[32], precursor and fragment tolerance were set to 4.5 ppm and 0.1 Da
620 respectively with carbamidomethylation of cysteine as a fixed modification and Nter-
621 Acetylation and methionine oxidation as variable modifications.Datasets were searched
622 using a target-decoy approach against a composite database composed of a target
623 database [Uniprot canonical and isoform reference proteome (16 January 2015) for a total
624 of 89,861 sequences + custom alternative proteome resulting from the in silico translation
625 of all human altORFs (available to download at
626 https://www.roucoulab.com/p/downloads)], and their reverse protein sequences from the
627 target database used as decoys. In order to separate alternative and reference proteins for
628 FDR analyses, PeptideShaker output files were extracted with target and decoy hits.
629 PSMs matching reference target or decoy proteins were separated from those matching
630 alternative targets or decoys as previously described [83,84]. PSMs that matched both
631 reference and alternative proteins were automatically moved to the reference database
632 group. PSMs were then ranked according to their PeptideShaker score and filtered at 1%
633 FDR separately. Validated PSMs were selected to group proteins using proteoQC R tool
634 [85] , and proteins were separately filtered again using a 1% FDR cut-off.
635 Only alternative proteins identified with at least one unique and specific peptide were
636 considered valid[59]. Any peptide matching both a canonical (annotated in Uniprot) and an
637 alternative protein was attributed to the canonical protein. For non-unique peptides, i.e.
638 peptides matching more than one alternative protein, the different accession IDs are
639 indicated in the MS files. For subsequent analyses (e.g. conservation, protein

signature…), only one protein is numbered in the total count of alternative proteins; we arbitrarily selected the alternative protein with the lowest accession ID.

Peptides matching proteins in a protein sequence database for common contaminants were rejected[86].

For spectral validation (Figure 6-figure supplement1, 2, 3, and 4), synthetic peptides were purchased from the peptide synthesis service at the *Université de Sherbrooke*. Peptides were solubilized in 10% acetonitrile, 1% formic acid and directly injected into a Q-Exactive mass spectrometer (Thermo Scientific) via an electro spray ionization source (Thermo Scientific). Spectra were acquired using Xcalibur 2.2 (RRID:SCR_014593) 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. Alternative protein sequence 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 corresponding to human MiD51/MIEF1/SMCR7L transcript variant 1 (NM_019008) was also cloned into pcDNA3.1 by Gibson assembly. In this construct, altMiD51 and MiD51 were tagged with Flag and HA tags, respectively. $MiD51^{GFP}$ and $altMiD51^{GFP}$ were also cloned into pcDNA3.1 by Gibson assembly. For $MiD51^{GFP}$, a LAP tag[32] was inserted between MiD51 and GFP. gBlocks were purchased from IDT. Human $altDDIT3^{mCherry}$ was cloned into pcDNA3.1 by Gibson assembly using coding sequence from transcript variant 1 (NM_004083.5) and mCherry coding sequence from pLenti-myc-GLUT4-mCherry (Addgene plasmid # 64049). Human $DDIT3^{GFP}$ was also cloned into pcDNA3.1 by Gibson assembly using CCDS8943 sequence.

For immunofluorescence, primary antibodies were diluted as follow: anti-Flag (Sigma, F1804) 1/1000, anti-TOM20 (Abcam, ab186734) 1/500. For western blots, primary antibodies were diluted as follow: anti-Flag (Sigma, F1804) 1/1000, anti-HA (BioLegend, 901515) 1/500, anti-actin (Sigma, A5441) 1/10000, anti-Drp1 (BD Transduction Laboratories, 611112) 1/500, anti-GFP (Santa Cruz Biotechnology, sc-9996) 1/10000, anti-mCherry (Abcam, ab125096) 1/2000.

685 **Cell culture, immunofluorescence, knockdown and western blots.** HeLa cells (ATCC
686 CRM-CCL-2, authenticated by STR profiling, RRID:CVCL_0030) cultures tested
687 negative for mycoplasma contamination (ATCC 30-1012K), transfections,
688 immunofluorescence, confocal analyses and western blots were carried out as previously
689 described[87]. Mitochondrial morphology was analyzed as previously described[88]. A
690 minimum of 100 cells were counted (n=3 or 300 cells for each experimental condition).
691 Three independent experiments were performed.
692 For Drp1 knockdown, 25,000 HeLa cells in 24-well plates were transfected with 25 nM
693 Drp1 SMARTpool: siGENOME siRNA (Dharmacon, M-012092-01-0005) or ON-
694 TARGET plus Non-targeting pool siRNAs (Dharmacon, D-001810-10-05) with
695 DharmaFECT 1 transfection reagent (Dharmacon, T-2001-02) according to the
696 manufacturer's protocol. After 24h, cells were transfected with pcDNA3.1 or altMiD51,
697 incubated for 24h, and processed for immunofluorescence or western blot. Colocalization
698 analyses were performed using the JACoP plugin (Just Another Co-localization Plugin) [50]
699 implemented in Image J software.
700
701 **Immunoprecipitations.** Immunoprecipitations experiments were conducted using GFP-
702 Trap (ChromoTek) protocol with minor modifications. Briefly, cells were lysed with co-
703 ip lysis buffer (0.5 % NP40, Tris-HCl 50 mM pH 7.5, NaCl 150 mM and two EDTA-free
704 Roche protease inhibitors per 50 mL of buffer). After 5 mins of lysis on ice, lysate was
705 sonicated twice at 11 % amplitude for 5 s with 3 minutes of cooling between sonication
706 cycles. Lysate was centrifuged, supernatant was isolated and protein content was assessed
707 using BCA assay (Pierce). GFP-Trap beads were conditioned with lysis buffer. 40 µL of

beads were added to 2 mg of proteins at a final concentration of 1 mg/mL. After overnight immunoprecipitation, beads were centrifuged at 5000 rpm for 5 minutes and supernatant was discarded. Beads were then washed three times with wash buffer (0.5 % NP40, Tris-HCl 50 mM pH 7.5, NaCl 200 mM and two EDTA-free Roche protease inhibitors per 50 mL of buffer) and supernatants were discarded. Immunoprecipitated proteins were eluted from beads by adding 40 µL of Laemmli buffer and boiling at 95 °C for 15 minutes. Eluate was split in halfs which were loaded onto 10 % SDS-PAGE gels to allow western blotting of GFP and mCherry tagged proteins. 40 µg of initial lysates were loaded into gels as inputs.

Mitochondrial localization, parameters and ROS production. Trypan blue quenching experiment was performed as previously described[89].

A flux analyzer (XF96 Extracellular Flux Analyzer; Seahorse Bioscience, Agilent technologies) was used to determine the mitochondrial function in HeLa cells overexpressing altMiD51^{Flag}. Cells were plated in a XF96 plate (Seahorse Biosciences) at 1×10^4 cells per well in Dulbecco's modified Eagle's medium supplemented with 10% FBS with antibiotics. After 24 hours, cells were transfected for 24 hours with an empty vector (pcDNA3.1) or with the same vector expressing altMiD51^{Flag} with GeneCellin tranfection reagent according to the manufacturer's instructions. Cells were equilibrated in XF assay media supplemented with 25 mM glucose and 1 mM pyruvate and were incubated at 37 °C in a CO_2-free incubator for 1 h. Baseline oxygen consumption rates (OCRs) of the cells were recorded with a mix/wait/measure times of 3/0/3 min respectively. Following these measurements, oligomycin (1 µM), FCCP (0.5 µM), and

731 antimycin A/rotenone (1 µM) were sequentially injected, with oxygen consumption rate
732 measurements recorded after each injection. Data were normalized to total protein in each
733 well. For normalization, cells were lysed in the 96-well XF plates using 15 µl/well of
734 RIPA lysis buffer (1% Triton X-100, 1% NaDeoxycholate, 0.1% SDS, 1 mM EDTA, 50
735 mM Tris-HCl pH 7.5). Protein concentration was measured using the BCA protein assay
736 reagent (Pierce, Waltham, MA, USA).
737 Reactive oxygen species (ROS) levels were measured using Cellular ROS/Superoxide
738 Detection Assay Kit (Abcam #139476). HeLa cells were seeded onto 96-well black/clear
739 bottom plates at a density of 6,000 cells per well with 4 replicates for each condition.
740 After 24 hours, cells were transfected for 24 hours with an empty vector (pcDNA3.1) or
741 with the same vector expressing altMiD51^{Flag} with GeneCellin according to the
742 manufacturer's instruction. Cells were untreated or incubated with the ROS inhibitor (N-
743 acetyl-L-cysteine) at 10 mM for 1 hour. Following this, the cells were washed twice with
744 the wash solution and then labeled for 1 hour with the Oxidative Stress Detection
745 Reagent (green) diluted 1:1000 in the wash solution with or without the positive control
746 ROS Inducer Pyocyanin at 100 µM. Fluorescence was monitored in real time. ROS
747 accumulation rate was measured between 1 to 3 hours following induction. After the
748 assay, total cellular protein content was measured using BCA protein assay reagent
749 (Pierce, Waltham, MA, USA) after lysis with RIPA buffer. Data were normalised for
750 initial fluorescence and protein concentration.
751 ATP synthesis was measured as previously described[90] in cells transfected for 24 hours
752 with an empty vector (pcDNA3.1) or with the same vector expressing altMiD51^{Flag}.
753

754 **Acknowledgements**

755 This research was supported by CIHR grants MOP-137056 and MOP-136962 to X.R;
756 MOP-299432 and MOP-324265 to C.L; a *Université de Sherbrooke* institutional research
757 grant made possible through a generous donation by Merck Sharp & Dohme to X.R; a
758 FRQNT team grant 2015-PR-181807 to C.L. and X.R; Canada Research Chairs in
759 Functional Proteomics and Discovery of New Proteins to X.R, in Evolutionary Cell and
760 Systems Biology to C.L and in Computational and Biological Complexity to A.O; A.A.C
761 is supported by a CIHR New Investigator Salary Award; M.S.S is a recipient of a *Fonds*
762 *de Recherche du Québec – Santé* Research Scholar Junior 2 Career Award; V.D is
763 supported in part by fellowships from *Région Nord-Pas de Calais* and PROTEO; A.A.C,
764 D.J.H, M.S.S and X.R are members of the *Fonds de Recherche du Québec Santé-*
765 supported *Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke*. We
766 thank the staff from the Centre for Computational Science at the *Université de*
767 *Sherbrooke*, Compute Canada and Compute Québec for access to the Mammouth
768 supercomputer.

769

770

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).

21. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17,** 567–578 (2016).

22. Sabath, N., Wagner, A. & Karlin, D. Evolution of viral proteins originated de novo by overprinting. *Mol. Biol. Evol.* **29,** 3767–80 (2012).

23. Iacono, M., Mignone, F. & Pesole, G. uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* **349,** 97–105 (2005).

24. Neafsey, D. E. & Galagan, J. E. Dual modes of natural selection on upstream open reading frames. *Mol. Biol. Evol.* **24,** 1744–51 (2007).

25. Smith, E. *et al.* Leaky ribosomal scanning in mammalian genomes: significance of histone H4 alternative translation in vivo. *Nucleic Acids Res.* **33,** 1298–1308 (2005).

26. Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.* **10,** 770 (2014).

27. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).

28. Michel, A. M. *et al.* GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.* **42,** D859-864 (2014).

29. Raj, A. *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5,** 1–24 (2016).

30. Miettinen, T. P. & Björklund, M. Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic Acids Res.* **43,** 1019–1034 (2015).

31. Weingarten-Gabbay, S. *et al.* Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science (80-.).* **351,** 1–24 (2016).

32. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163,** 712–723 (2015).

33. Tong, J., Taylor, P. & Moran, M. F. Proteomic analysis of the epidermal growth factor receptor (EGFR) interactome and post-translational modifications associated with receptor endocytosis in response to EGF and stress. *Mol. Cell. Proteomics* **13,** 1644–1658 (2014).

34. Sharma, K. *et al.* Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Rep.* **8,** 1583–1594 (2014).

35. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. data* **1,** 140031 (2014).

36. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43,** D213-221 (2014).

37. Vanderperre, B. *et al.* Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* **8,** e70698 (2013).

38. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 183–212 (2000).

39. Laity, J. H., Lee, B. M. & Wright, P. E. Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* **11,** 39–46 (2001).

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.

Molecular symbiosis of CHOP and C/EBP beta isoform LIP contributes to endoplasmic reticulum stress-induced apoptosis. *Mol. Cell. Biol.* **30,** 3722–31 (2010).

50. Bolte, S. & Cordelières, F. P. A guided tour into subcellular colocalization analysis in light microscopy. *J. Microsc.* **224,** 213–32 (2006).

51. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* (2017). doi:10.1038/nrm.2017.58

52. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* (2016). doi:10.15252/embj.201592759

53. Jousse, C. *et al.* Inhibition of CHOP translation by a peptide encoded by an open reading frame localized in the chop 5'UTR. *Nucleic Acids Res.* **29,** 4341–51 (2001).

54. Young, S. K., Palam, L. R., Wu, C., Sachs, M. S. & Wek, R. C. Ribosome Elongation Stall Directs Gene-specific Translation in the Integrated Stress Response. *J. Biol. Chem.* **291,** 6546–6558 (2016).

55. Lee, Y. C. G. & Reinhardt, J. A. Widespread Polymorphism in the Positions of Stop Codons in Drosophila melanogaster. *Genome Biol. Evol.* **4,** 533–549 (2012).

56. Andreatta, M. E. *et al.* The Recent De Novo Origin of Protein C-Termini. *Genome Biol. Evol.* **7,** 1686–701 (2015).

57. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res.* **19,** 1752–9 (2009).

58. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a

model of frequent de novo evolution. *BMC Genomics* **14,** 117 (2013).

59. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33,** 22–24 (2015).

60. Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10,** M111.007690 (2011).

61. 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).

62. 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).

63. 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).

64. 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).

65. 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).

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

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

68. 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).

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

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

71. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

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

73. 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).

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

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

76. 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).

77. 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).

78. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467 (2004).

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

80. 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).

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

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

83. Menschaert, G. & Fenyö, D. Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrom. Rev.* **36,** 584–599 (2015).

84. Woo, S. *et al.* Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics* **14,** 2719–30 (2014).

85. Gatto, L., Breckels, L. M., Naake, T. & Gibb, S. Visualization of proteomics data

using R and Bioconductor. *Proteomics* **15,** 1375–1389 (2015).

86. 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).

87. 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).

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

89. 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

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

992 **FIGURE LEGENDS**

993

994 **Figure 1. Annotation of human altORFs.**

995 (**a**) AltORF nomenclature. AltORFs partially overlapping the CDS must be in a different
996 reading frame. (**b**) Pipeline for the identification of altORFs. (**c**) Size distribution of
997 alternative (empty bars, vertical and horizontal axes) and reference (grey bars, secondary
998 horizontal and vertical axes) proteins. Arrows indicate the median size. The median
999 alternative protein length is 45 amino acids (AA) compared to 460 for the reference
1000 proteins. (**d**) Distribution of altORFs in the human hg38 transcriptome. (**e, f**) Number of
1001 total altORFs (**e**) or number of altORFs/10kbs (**f**) in hg38 compared to shuffled hg38.
1002 Means and standard deviations for 100 replicates obtained by sequence shuffling are
1003 shown. Statistical significance was determined by using one sample t-test with two-tailed
1004 *p*-values. **** $p<0.0001$. (**g**) Percentage of altORFs with an optimal Kozak motif. The
1005 total number of altORFs with an optimal Kozak motif is also indicated at the top.

1006

1007 **Figure 1-figure supplement 1. 10% of altORFs are present in different classes of**
1008 **repeats.**

1009 While more than half of the human genome is composed of repeated sequences, only
1010 9.83% or 18,003 altORFs are located inside these repeats (**a**), compared to 2,45% or
1011 1,677 CDSs (**b**). AltORFs and CDSs are detected in non-LTR retrotransposons (LINEs,
1012 SINEs, SINE-VNTR-Alus), LTR repeats, DNA repeats, satellites and other repeats.
1013 Proportions were determined using RepeatMasker (version 3.3.0).

1014

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 539,134 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%).

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 between *H. sapiens* and other species (pairwise study).

Figure 3. 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 relative to the total number of reference-alternative protein pairs are indicated on the right (see Table 2 for details).

Figure 4. 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). We restricted the analysis to altORF-CDS pairs that were co-conserved from humans to zebrafish. The plot contains 889 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 signaling conservation DQ 0.95, and blue signaling accelerated evolution DQ 0.05) with 317 of 889 altORFs (35.7%). A random distribution would have implied a total of 10% (or 89) of altORFs in the extreme values. This suggests that 25.7% (35.7%-10%) of these 889 altORFs undergo specific selection different from random regions in their CDSs with a similar length distribution. This percentage is very similar to the 26.2% obtained from an analysis of altORFs without restriction based on co-conservation in vertebrates (see Figure 4-figure

supplement 1), a total which would imply that there are about 4,458 altORFs fully nested in CDSs undergoing conserved or accelerated evolution relative to their CDSs.

Figure 4-figure supplement 1. 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 third codon positions 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 5. 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 6. 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[32], 10,000 human proteins[35], human phosphoproteome[34], EGFR interactome[33]. Numbers are indicates above each column. (**c**) Percentage of altORFs encoding alternative proteins detected by MS-based proteomics. The total number of altORFs is indicated at the top. Localization “Unknown” indicates that the detected peptides can match more than one alternative protein. Localization ">1" indicates that the altORF can have more than one localization in different RNA isoforms.

Figure 6-figure supplement 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.

Figure 6-figure supplement 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.

Figure 6-figure supplement 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.

Figure 6-figure supplement 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.

Figure 7. The alternative phosphoproteome in mitosis and EGF-treated cells.

Heatmap showing relative levels of spectral counts for phosphorylated peptides following

the indicated treatment[34]. 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 above 10 are shown.

Figure 7-figure supplement 1: Example of a phosphorylated peptide in mitosis - alternative protein AltLINC01420^{nc} (LOC550643, IP_305449.1).

(**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.

Figure 8. Number of alternative proteins detected by ribosome profiling and mass spectrometry.

The expression of 467 alternative proteins was detected by both ribosome profiling (translation initiation sites, TIS) and mass spectrometry (MS).

Figure 9. 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 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 10. 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 11. 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 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 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.

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 (100 cells for each independent experiment). ***p<0.0005 (Fisher's exact test) for the three morphologies between altMiD51(WT) and the other experimental conditions.

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 mm. (**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 three independent experiments (8 technical replicates each). 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 mm.

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). Representative experiment of three independent biological replicates.

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 (100 cells for each independent experiment). ***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. (**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 mm. (**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 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_004083.5). +2 and +1 indicate reading frames. AltDDIT3 amino acid sequence is also shown. (**b**) Confocal microscopy analyses of HeLa cells co-transfected with altDDIT3^{GFP} (green channel) and DDIT3^{mCherry} (red channel). Scale bar, 10 µm. (**c, d**) Colocalization analysis of the images shown in (**b**) performed using the JACoP plugin (Just Another Co-localization Plugin) implemented in Image J software (two independent

biological replicates). (**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 (**b**) 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 (two independent experiments).

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). Representative experiment of two independent biological replicates.

Figure 15–figure supplement 2. Colocalization of altDDIT3 with DDIT3.

Scatter plots of Pearson's Correlation Coefficient and Manders' Correlation Coefficient

after Costes' automatic threshold (p-value < 0.001, based on 1000 rounds of Costes' randomization colocalization analysis). M1 is the proportion of altDDIT3^{GFP} signal overlapping DDIT3^{mCherry} signal and M2 is the proportion of DDIT3^{mCherry} signal overlapping altDDIT3^{GFP}. Error bars represent the mean +/- SD of three independent experiments (28 cells).

SUPPLEMENTARY FILES

Supplementary file 1: 12,616 alternative proteins and 26,531 reference proteins with translation initiation sites detected by ribosome profiling after re-analysis of large scale studies. Sheet 1: general information. Sheet 2: list of alternative proteins; sheet 3: pie chart of corresponding altORFs localization. Sheet 4: Sheet 2: list of reference proteins

Supplementary file 2: 4,872 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: 383 alternative proteins detected by mass spectrometry in the interactome of 118 zinc finger proteins. Sheet 1: MS identification parameters; sheet 2: raw MS output; sheet 3: list of detected alternative proteins.

Supplementary file 6: high-confidence list of predicted functional and co-operating alternative proteins based on co-conservation and expression analyses. Sheet 1: high

confidence list in mammals; sheet 2: high confidence list in in vertebrates.

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	68,264	54,498	539,134	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: 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/54,498)	% 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	47.20	44.27	50.39	<1e-06
M. musculus	25,930	22,304	24,862	157,261	106,987	95.88	40.09	68.031	69.39	<1e-06
B. taurus	25,868	16,887	24,426	157,323	99,369	94.42	30.98	63.16	64.67	<1e-06
X. tropicalis	2,470	12,458	1,974	180,721	95,499	79.91	22.85	52.84	57.81	<1e-06
D. rerio	2,023	12,791	1,203	181,168	94,426	59.46	23.47	52.12	57.29	<1e-06
D. melanogaster	115	4,881	51	183,076	34,352	44.34	8.95	18.76	38.26	<1e-06
C. elegans	34	3,954	8	183,157	26,839	23.52	7.25	14.65	50.00	0.02
S. cerevisiae	6	1,854	2	183,185	10,935	33.33	3.40	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.

Table 3: alternative zinc finger proteins detected by mass spectrometry (MS) and ribosome profiling (RP)

Formatted: Numbering: Continuous

Alternative protein accession	Detection method[1]	Gene	Amino acid sequence	AltORF localization
IP_238718.1	MS	RP11	MLVEVACSSCRSLLHKGAGASEDGAALEPAHTGGKENGATT	nc
IP_278905.1	RP	ZNF761	MSVARPLVGSHILYAIIDFILERNLISVMSVARTLVRSHPLYAT IDFILERNLTSVMSVARPLVRSQTLHAIVDFILEKNKCNECGE VFNQQAHLAGHHRIHTGEKP	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'UTR
IP_275012.1	MS	ZNF780A	MKPCECTECGKTFSCSSNIVQHVKIHTGEKRYNVRNMGKHLL WMISCLNIRKFRIVRNFVTIRSVDKPSLCTKNLLNTRELILMRN LVNIKECVKNFHHGLGFAQLLSIHTSEKSLSVRNVGRFIATLN TLEFGEDNSCEKVFE	3'UTR
IP_270595.1[2]	MS	ZNF440	MHSVERPYKCKICGRGFYSAKSFQIHEKSYTGEKPYECKQCG KAFVSFTSFRYHERTHTGENPYECKQFGKAFRSVKNLRFHKR THTGEKPCECKKCRKAFHNFSSLQIHERMHRGEKLCECKHCG KAFISAKIL	CDS
IP_270643.1[2]	MS	ZNF763	MKKLTLERNPINACHVVKPSIFPVPFSIMKGLTLERNPMSVSV GKPSDVPHTFEGMVGLTGEKPYECKECGKAFRSASHLQIHER TQTHIRIHSGERPYKCKTCGKGFYSPTSFQRHEKTHTAEKPYE CKQCGKAFSSSSSFWYHERTHTGEKPYECKQCGKAFRSASIQ MHAGTHPEEKPYECKQCGKAFRSAPHLRIHGRTHTGEKPYEC KECGKAFRSAKNLRIHERTQTHVRMHSVERPYKCKICGKGFY SAKSFQIPEKSYTGEKPYECKQCGKAFISFTSFR	3'UTR
IP_270597.1[3]	MS	ZNF440	MKNLTLERNPMSVSNVGKPLFPSLPFDIMKGLTLERTPMSVS NLGKPSDLSKIFDFIKGHTLERNPVNVRNVEKHSIISLLCKYM KGCTEERSSVNVSIVGKHSYLPRSFEYMQEHTMERNPMNVK NAEKHSACLLPFIDMKRLTLEGNTMNASNVAKLSLLPVLFNI MKEHTREKPYQCKQCAKAFISSTSFQYHERTHMGEKPYECM PSGKAFISSSSLQYHERTHTGEKPYEYKQCGKAFRSASHLQM HGRTHTGEKPYECKQYGKAFRPDKIL	3'UTR
IP_270609.1[3]	MS	ZNF439	MNVSNVAKAFTSSSSFQYHERTHTGEKPYQCKQCGKAVRSA SRLQMHGSTHTWQKLYECKQYGKAFRSARIL	3'UTR
IP_270663.1[3]	MS	ZNF844	MHGRTHTQEKPYECKQCGKAFIFSTSFRYHERTHTGEKPYEC KQCGKAFRSATQLQMHRKIHTGEKPYECKQCGKAYRSVSQL LVHERTHTVEQPYEYKQYGKAFRFAKNLQIQTMNVNN	CDS
IP_270665.1[3]	MS	ZNF844	MHRKIHTGEKPYECKQCGKAYRSVSQLLVHERTHTVEQPYE YKQYGKAFRFAKNLQIQTMNVNN	CDS
IP_270668.1[3]	MS	ZNF844	MSSTAFQYHEKTHTREKHYECKQCGKAFISSGSLRYHERTHT GEKPYECKQCGKAFRSATQLQMHRKIHTGEKPYECKQCGKA YRSVSQLLVHERTHTVEQPYEYKQYGKAFRFAKNLQIQTMN	3'UTR

			VNN	
IP_138139.1	MS	ZNF322	MLSPSRCKRIHTGEQLFKCLQCQLCCRQYEHLIGPQKTHPGE KPQQV	3'UTR
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 RTHTRGMSYEGDECRGL	CDS
IP_275016.1	RP	ZNF780A	MNVRSVGKALIVVHTLFSIRKFIPMRNLLYVGNVRWPLDIIAN LLNILEFILVTSHLNVKTVGRPSIVAQALFNIRVFTLVRSPMNV RSVGRLLDFTYNFPNIRKLTQVKNHLNVRNVGNSFVVVQILI NIEVFILERNPLNVRNVGKPFDFICTLFDIRNCILVRNPLNVRS VGKPFDFICNLFDIRNCILVRNPLNVRNVERFLVFPPSLIAIRTF TQVRRHLECKECGKSFNRVSNHVQHQSIRAGVKPCECKGCG KGFICGSNVIQHQKIHSSEKLFVCKEWRTTFRYHYHLFNITKF TLVKNPLNVKNVERPSVF	CDS or 3'UTR
IP_278870.1	RP	ZNF845	MNVARFLIEKQNLHVIIEFILERNIRNMKNVTKFTVVNQVLKD RRIHTGEKAYKCKSL	CDS
IP_278888.1	RP	ZNF765	MSVARPSAGRHPLHTIIDFILDRNLTNVKIVMKLSVSNQTLKD IGEFILERNYTCNECGKTFNQELTLTCHRRLHSGEKPYKYEEL DKAYNFKSNLEIHQKIRTEENLTSVMSVARP	CDS
IP_278918.1	RP	ZNF813	MNVARVLIGKHTLHVIIDFILERNLTSVMNVARFLIEKHTLHIII DFILEINLTSVMNVARFLIKKHTLHVTIDFILERNLTSVMNVAR FLIKKQTLHVIIDFILERNLTSLMSVAKLLIEKQSLHIIIQFILER NKCNECGKTFCHNSVLVIHKNSYWRETSVMNVAKFLINKHT FHVIIDFIVERNLRNVKHVTKFTVANRASKDRRIHTGEKAYK GEEYHRVFSHKSNLERHKINHTAEKP	CDS
IP_280349.1	RP	ZNF587	MNAVNVGNHFFPALRFMFIKEFILDKSLISAVNVENPFLNVPV SLNTGEFTLEKGLMNAPNVEKHFSEALPSFIIRVHTGERPYEC SEYGKSFAEASRLVKHRRVHTGERPYECCQCGKHQNVCCPR S	CDS
IP_280385.1	RP	ZNF417	MNAMNVGNHFFPALRFMFIKEFILDKSLISAVNVENPLLNVP VSLNTGEFTLEKGLMNVPNVEKHFSEALPSFIIRVHTGERPYE CSEYGKSFAETSRLIKHRRVHTGERPYECCQSGKHQNVCSPW S	CDS

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

[2] These two proteins were not detected with unique peptides but with shared peptides. One protein only was counted in subsequent analyses.

[3] These five proteins were not detected with unique peptides but with shared peptides. One protein only was counted in subsequent analyses.

1 **Table 4: Examples of proteins encoded in the same gene and functionally interacting**

Gene	Polypeptides[1]	Reference	altORF localization	altORF size aa	Conservation	Summary of functional relationship with the annotated protein
CDKN2A, INK4	Cyclin-dependent kinase inhibitor 2A or p16-INK4 (P42771), and p19ARF (Q8N726)	(61)	5'UTR	169	Unknown	the unitary inheritance of p16INK4a and p19ARF may underlie their dual requirement in cell cycle control.
GNAS, XLalphas	Guanine nucleotide-binding protein G(s) subunit alpha isoforms XL□s (Q5JWF2) and Alex (P84996)	(62)	5'UTR	+700	Human, mouse, rat	Both subunits transduce receptor signals into stimulation of adenylyl cyclase.
ATXN1	Ataxin-1 (P54253) and altAtaxin-1	(63)	CDS	185	Unknown	Direct interaction
Adora2A	A2A adenosine receptor (P30543) and uORF5	(64)	5'UTR	134	Human, chimpanzee, rat, mouse	A2AR stimulation increases the level of the uORF5 protein via post-transcriptional regulation.

AGTR1	Angiotensin type 1a receptor (P25095) and PEP7	(65)	5'UTR	7	Highly conserved across mammalian species	Inhibits non-G protein-coupled signalling of angiotensin II, without altering the classical G protein-coupled pathway activated by the ligand.

[1]The UniProtKB accession is indicated when available.

Figure 1

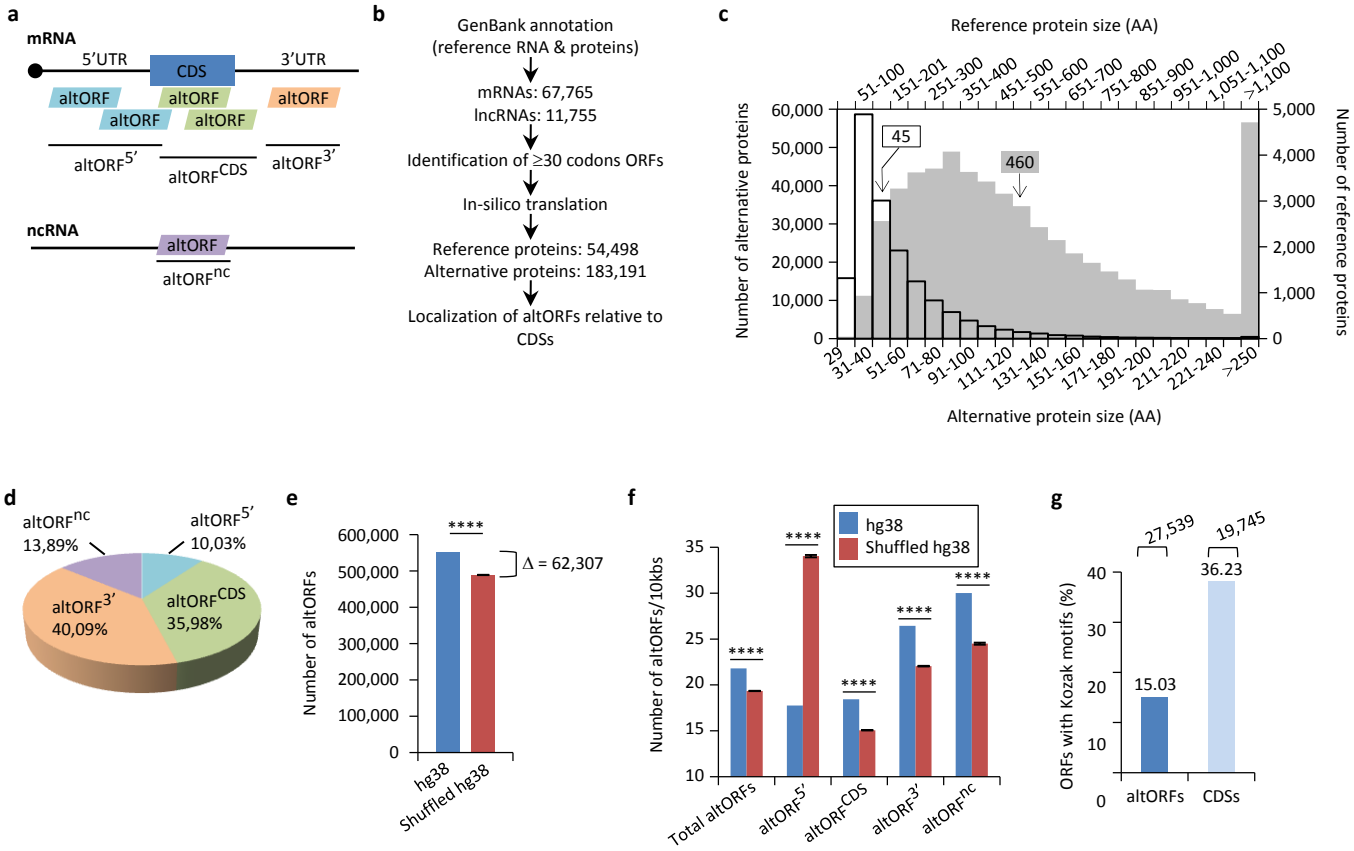


Figure 1. Annotation of human altORFs.
(**a**) AltORF nomenclature. AltORFs partially overlapping the CDS must be in a different reading frame. (**b**) Pipeline for the identification of altORFs. (**c**) Size distribution of alternative (empty bars, vertical and horizontal axes) and reference (grey bars, 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.

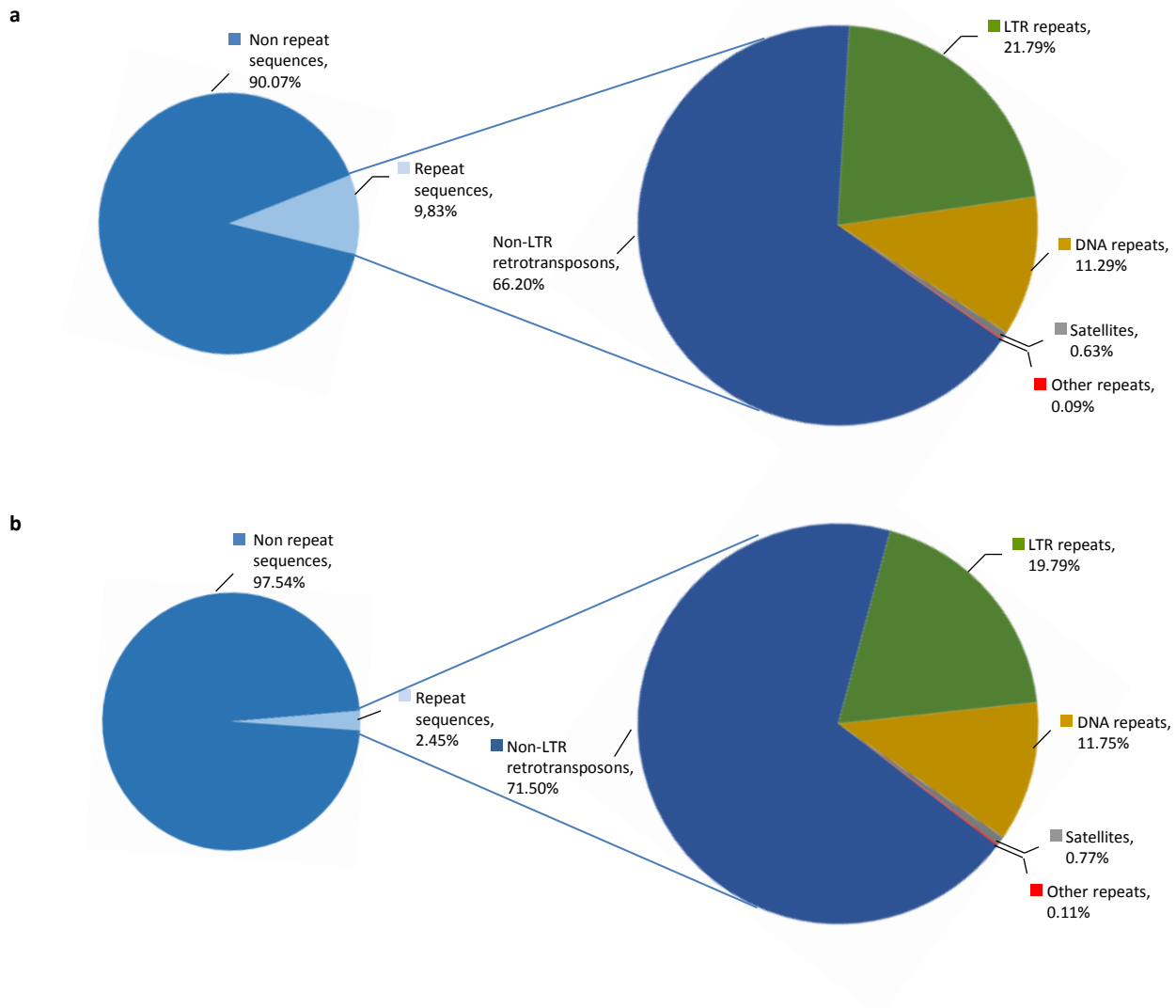


Figure 1-figure supplement 1: 10% of altORFs are present in different classes of repeats.

While more than half of the human genome is composed of repeated sequences, only 9.83% or 18,003 altORFs are located inside these repeats (**a**), compared to 2,45% or 1,677 CDSs (**b**). AltORFs and CDSs are detected in non-LTR retrotransposons (LINEs, SINEs, SINE-VNTR-Alus), LTR repeats, DNA repeats, satellites and other repeats. Proportions were determined using RepeatMasker (version 3.3.0).

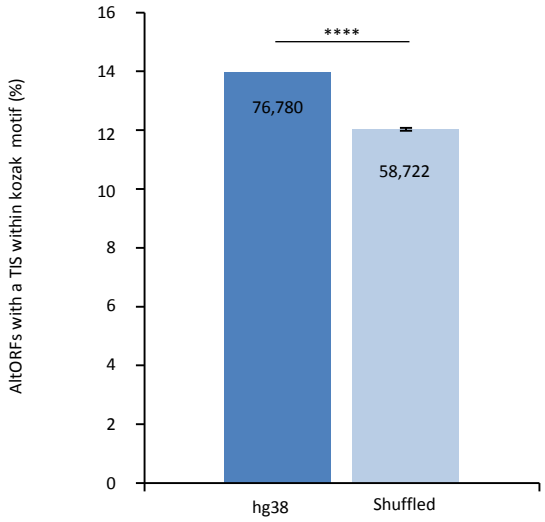


Figure 1-figure supplement 2: The proportion of altORFs with a translation initiation site (TIS) with a Kozak motif in hg38 is significantly different from 100 shuffled hg38 transcriptomes.

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 539,134 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%).

Figure 2

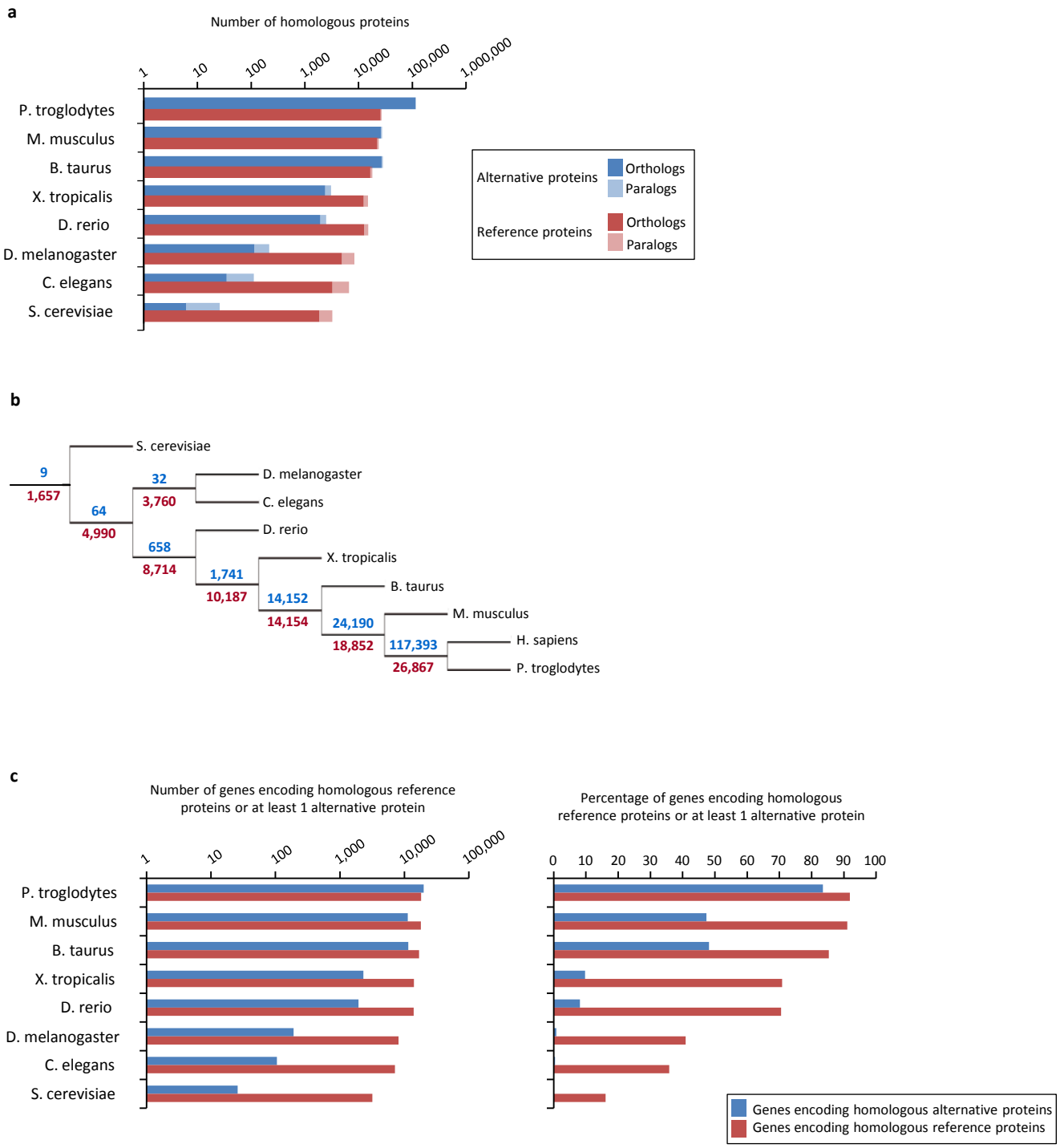


Figure 2. Conservation of alternative and reference proteins across different species. (**a**) Number of orthologous and paralogous alternative and reference proteins between *H. sapiens* and other species (pairwise study). (**b**) Phylogenetic tree: conservation of alternative (blue) and reference (red)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

Figure 3

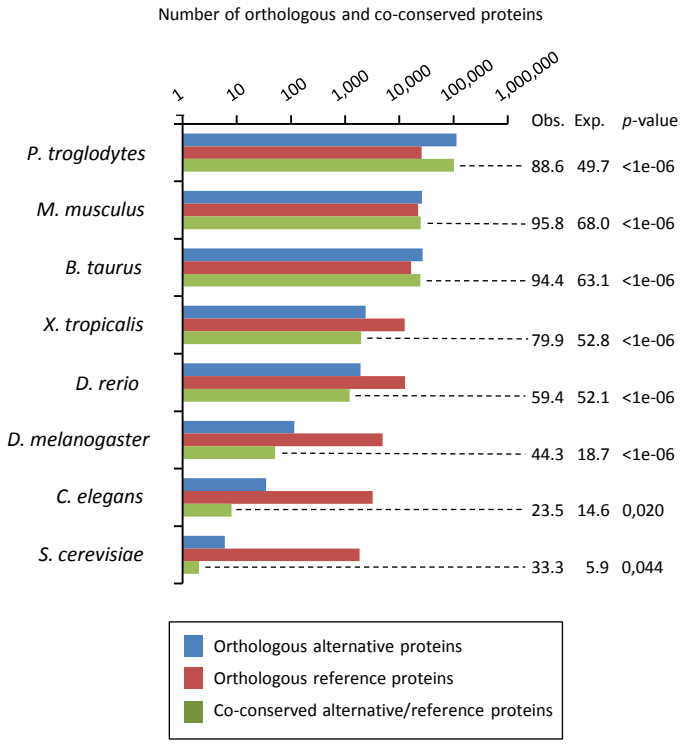


Figure 3. Number of orthologous and co-conserved alternative and reference proteins between *H. sapiens* and other species (pairwise). For the co-conservation analyses, the percentage of observed (Obs.), expected (Exp.) and corresponding *p*-values relative to the total number of reference-alternative protein pairs are indicated on the right (see Table 2 for details).

Figure 4

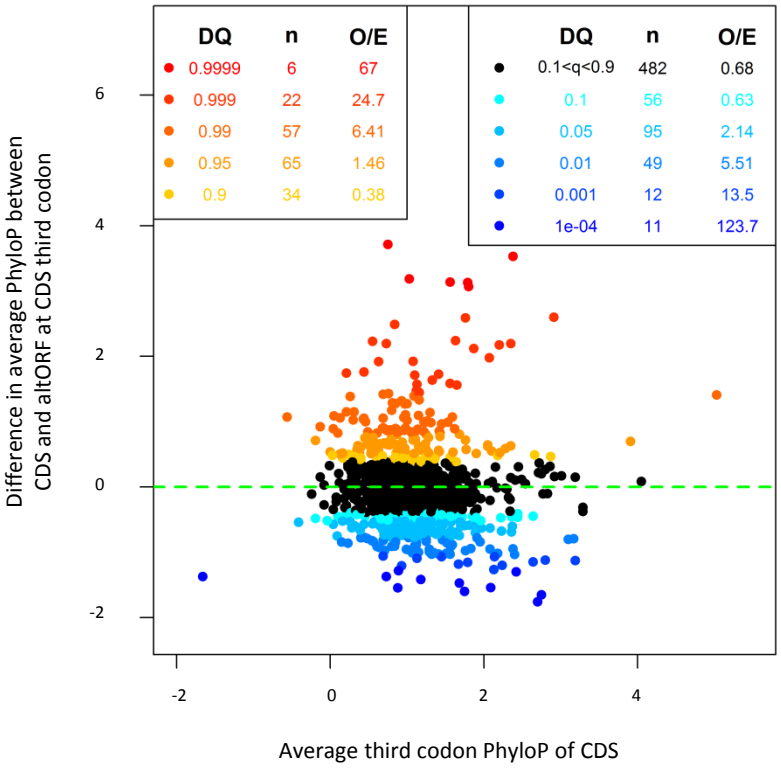


Figure 4. AltORFs completely nested within CDSs show more extreme PhyloP values (more conserved or faster evolving) than their CDSs. Differences between altORF and CDS PhyloP scores (altORF PhyloP – CDS PhyloP, *y*-axis) are plotted against PhyloPs for their respective CDSs (*x*-axis). We restricted the analysis to altORF-CDS pairs that were co-conserved from humans to zebrafish. The plot contains 889 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 signaling conservation DQ 0.95, and blue signaling accelerated evolution DQ 0.05) with 317 of 889 altORFs (35.7%). A random distribution would have implied a total of 10% (or 89) of altORFs in the extreme values. This suggests that 25.7% (35.7%-10%) of these 889 altORFs undergo specific selection different from random regions in their CDSs with a similar length distribution. This percentage is very similar to the 26.2% obtained from an analysis of altORFs without restriction based on co-conservation in vertebrates (see Figure 4-figure supplement 1), a total which would imply that there are about 4,458 altORFs fully nested in CDSs undergoing conserved or accelerated evolution relative to their CDSs.

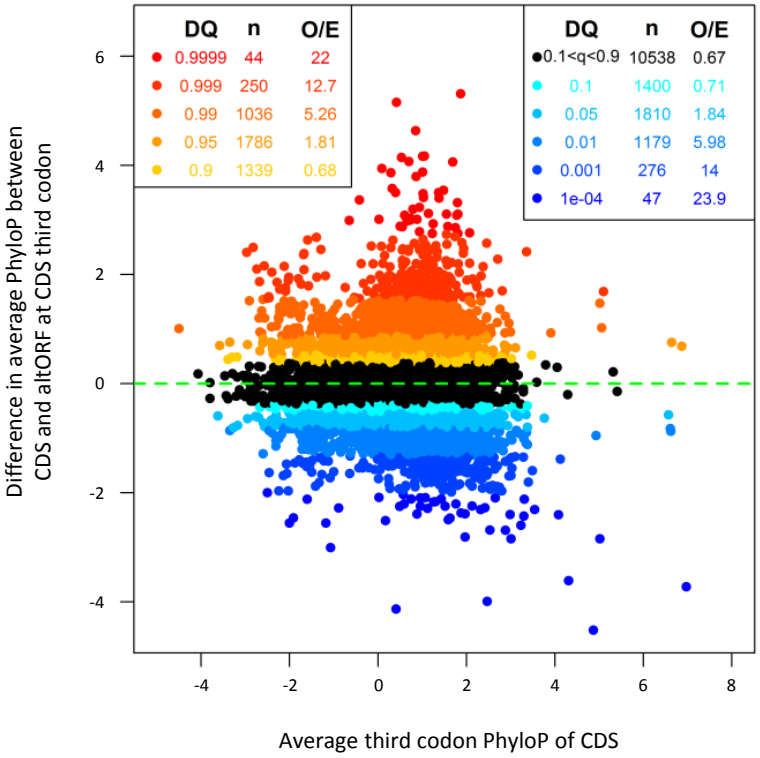


Figure 4-figure supplement 1. AltORFs completely nested within CDSs show more extreme PhyloP values (more conserved or faster evolving) than their CDSs. Differences between altORF and CDS PhyloP scores (altORF PhyloP – CDS PhyloP, *y*-axis) are plotted against PhyloPs for their respective CDSs (*x*-axis). 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 third codon positions 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 5

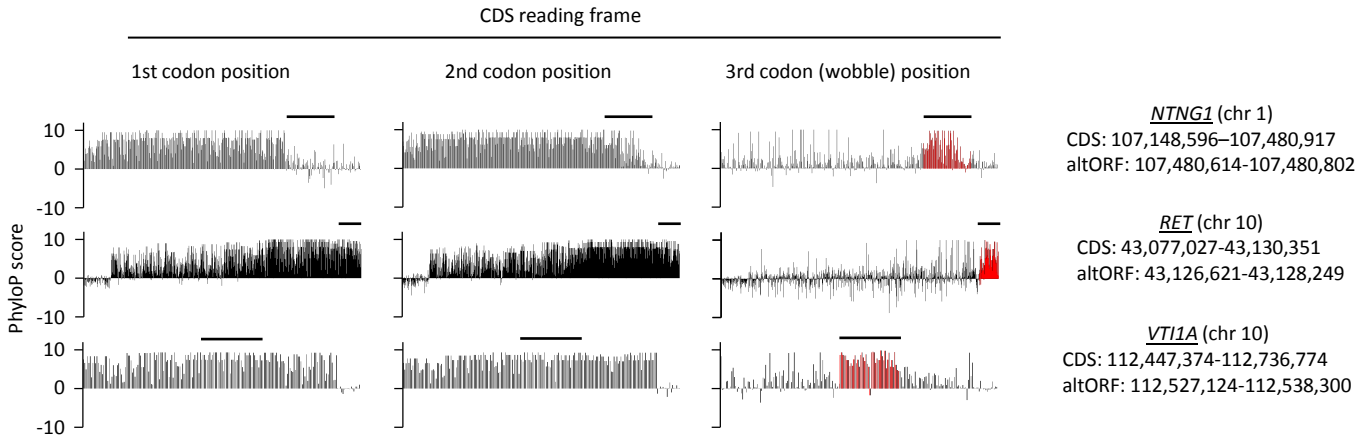


Figure 5. First, second, and third codon nucleotide PhyloP scores for 100 vertebrate species for the CDSs of the *NTNG1*, *RET* and *VTI1A* genes. Chromosomal coordinates for the different CDSs and altORFs are indicated on the right. The regions highlighted in red indicate the presence of an altORF characterized 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 6

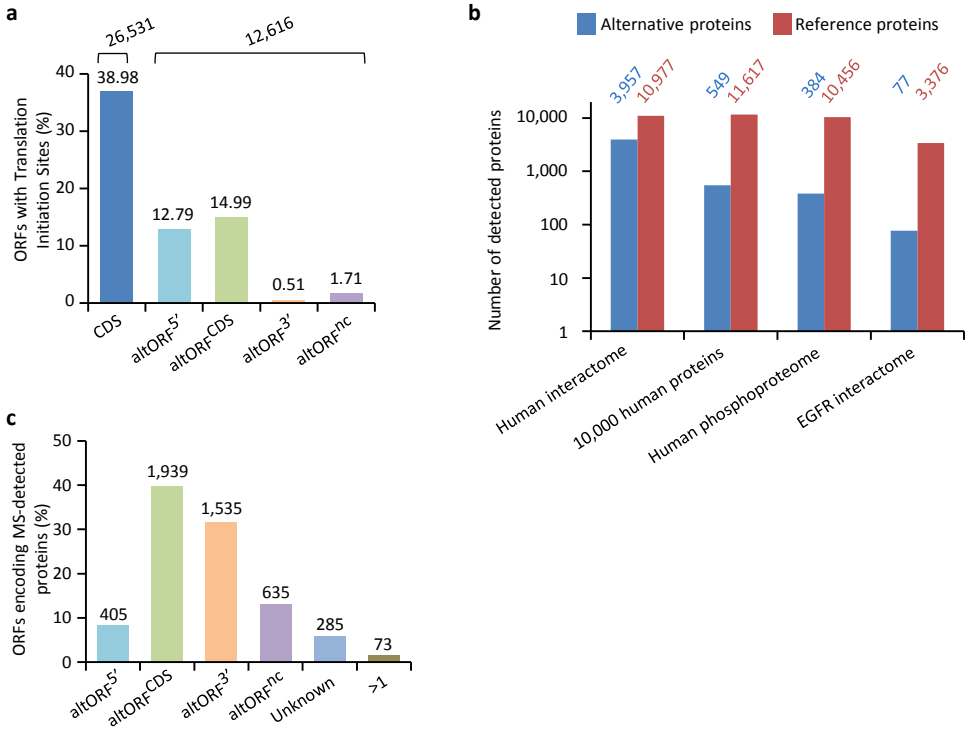


Figure 6. Expression of human altORFs.
(**a**) Percentage of CDSs and altORFs with detected TISs by ribosomal profiling and footprinting of human cells[23]. The total number of CDSs and altORFs with a detected TIS is indicated at the top. (**b**) Alternative and reference proteins detected in three large proteomic datasets: human interactome[32], 10,000 human proteins[35], human phosphoproteome[34], EGFR interactome[33]. Numbers are indicates above each column. (**c**) Percentage of altORFs encoding alternative proteins detected by MS-based proteomics. The total number of altORFs is indicated at the top. Localization “Unknown” indicates that the detected peptides can match more than one alternative protein. Localization ">1" indicates that the altORF can have more than one localization in different RNA isoforms.

Figure 7

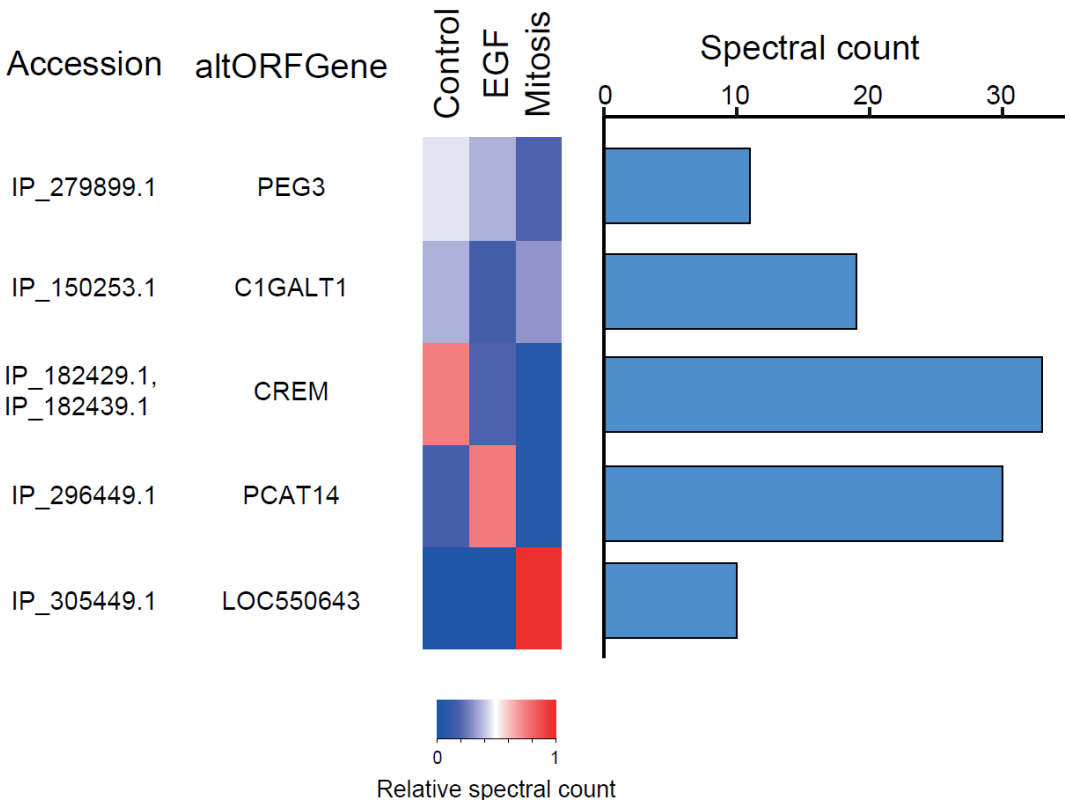


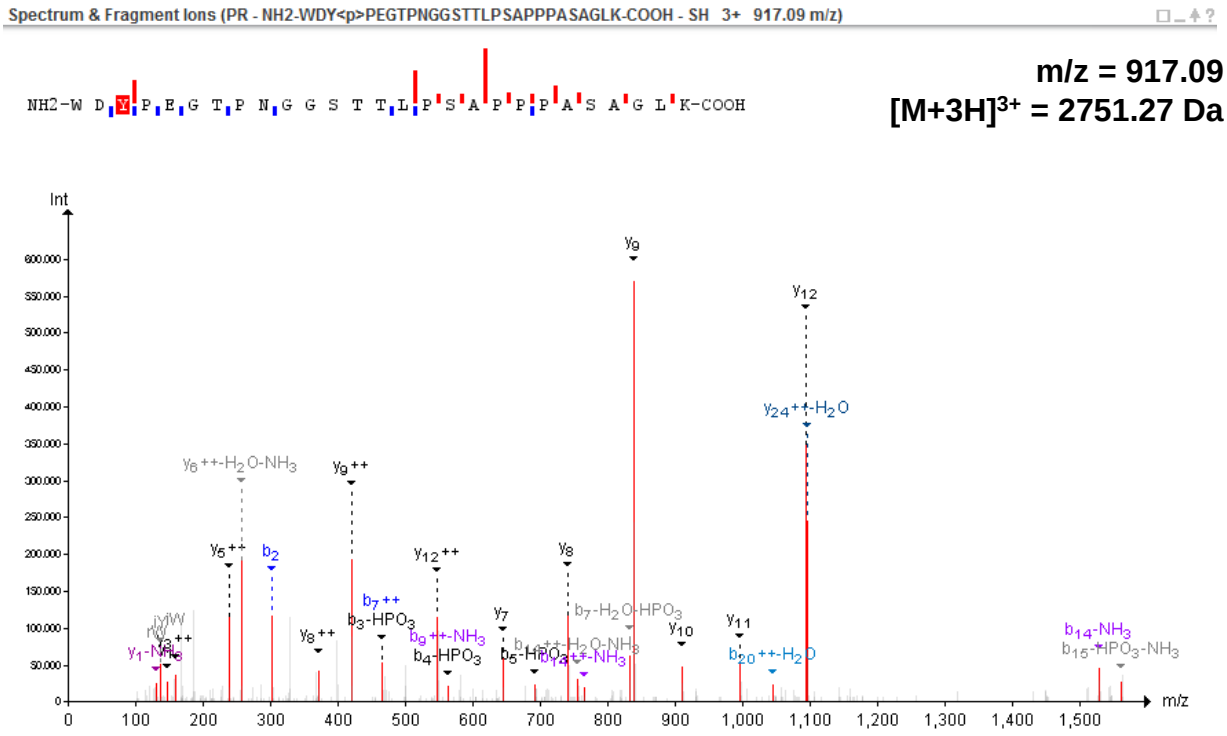
Figure 7. The alternative phosphoproteome in mitosis and EGF-treated cells. Heatmap showing relative levels of spectral counts for phosphorylated peptides following the indicated treatment[34]. For each condition, heatmap colors show the percentage of spectral count on total MS/MS phosphopeptide spectra. Blue bars on the right represent the number of MS/MS spectra; only proteins with spectral counts above 10 are shown.

a

AltLINC01420[nc]

MGDQPCASGRSTLPPGNAREAKPPKKRCLLAPR**WDYPEGTPNGGSTTLPSAPPPASAGLK**SHPPPPEK

b



c

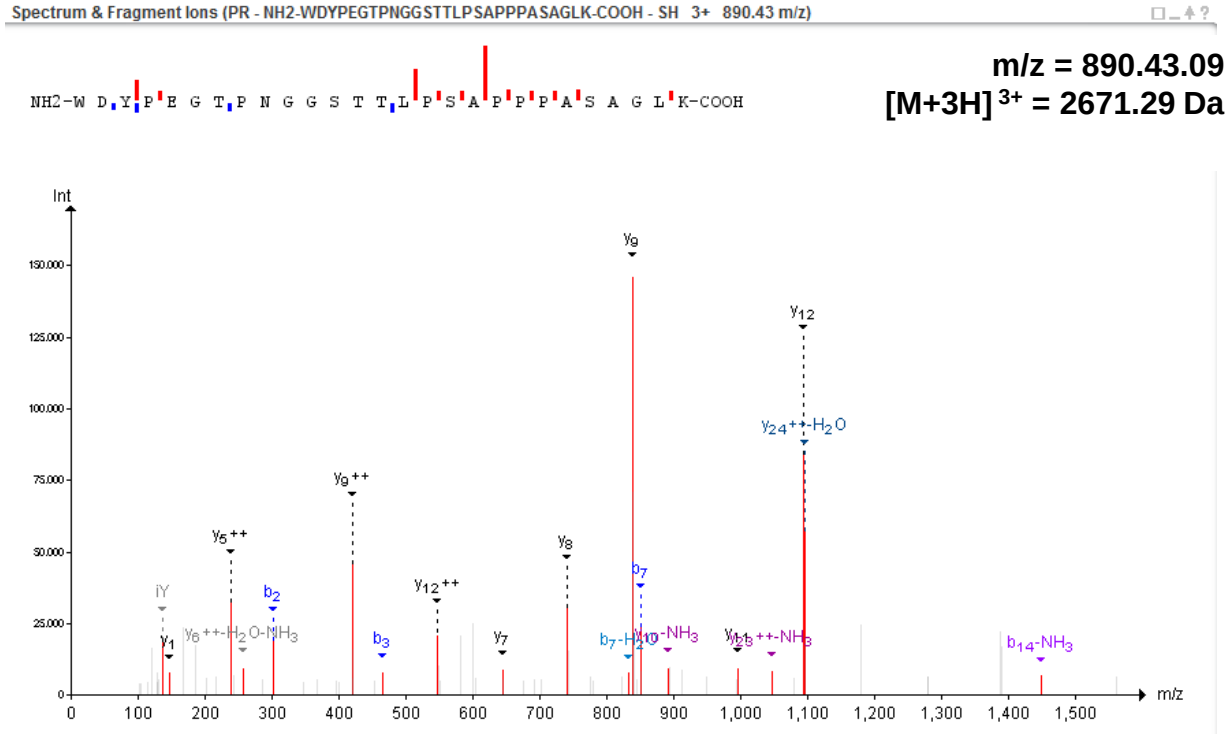


Figure 7-figure supplement 1: Example of a phosphorylated peptide in mitosis - alternative protein AltLINC01420^{nc} (LOC550643, IP_305449.1).
(**a**) AltLINC01420^{nc} amino acid sequence with detected peptides underlined and phosphorylated peptide in bold (73,9% sequence coverage). (**b**) MS/MS spectrum 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.

Figure 8

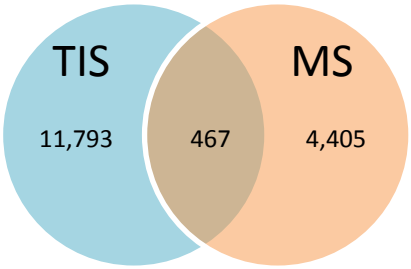


Figure 8. Number of alternative proteins detected by ribosome profiling and mass spectrometry.
The expression of 467 alternative proteins was detected by both ribosome profiling (translation initiation sites, TIS) and mass spectrometry (MS).

Figure 9

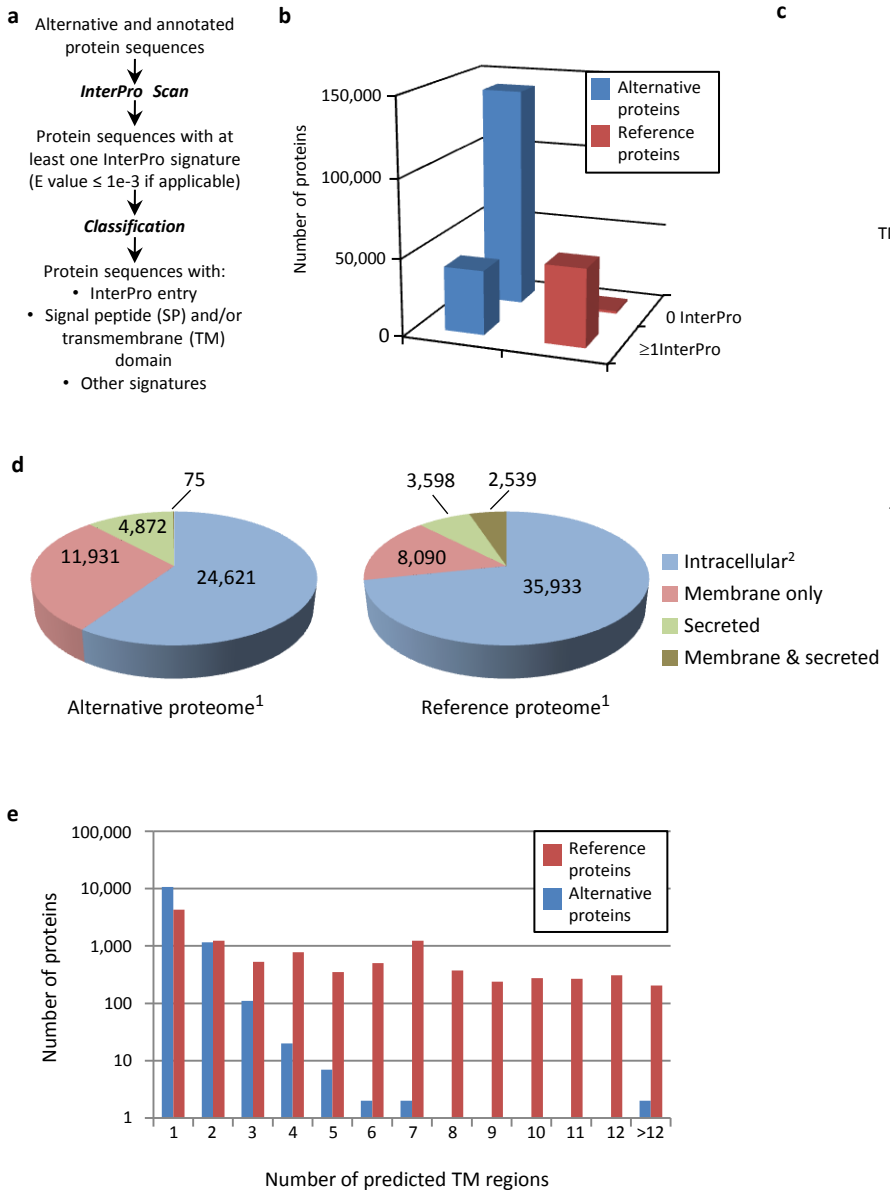


Figure 9. Human alternative proteome sequence analysis and classification using InterProScan.

(**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.

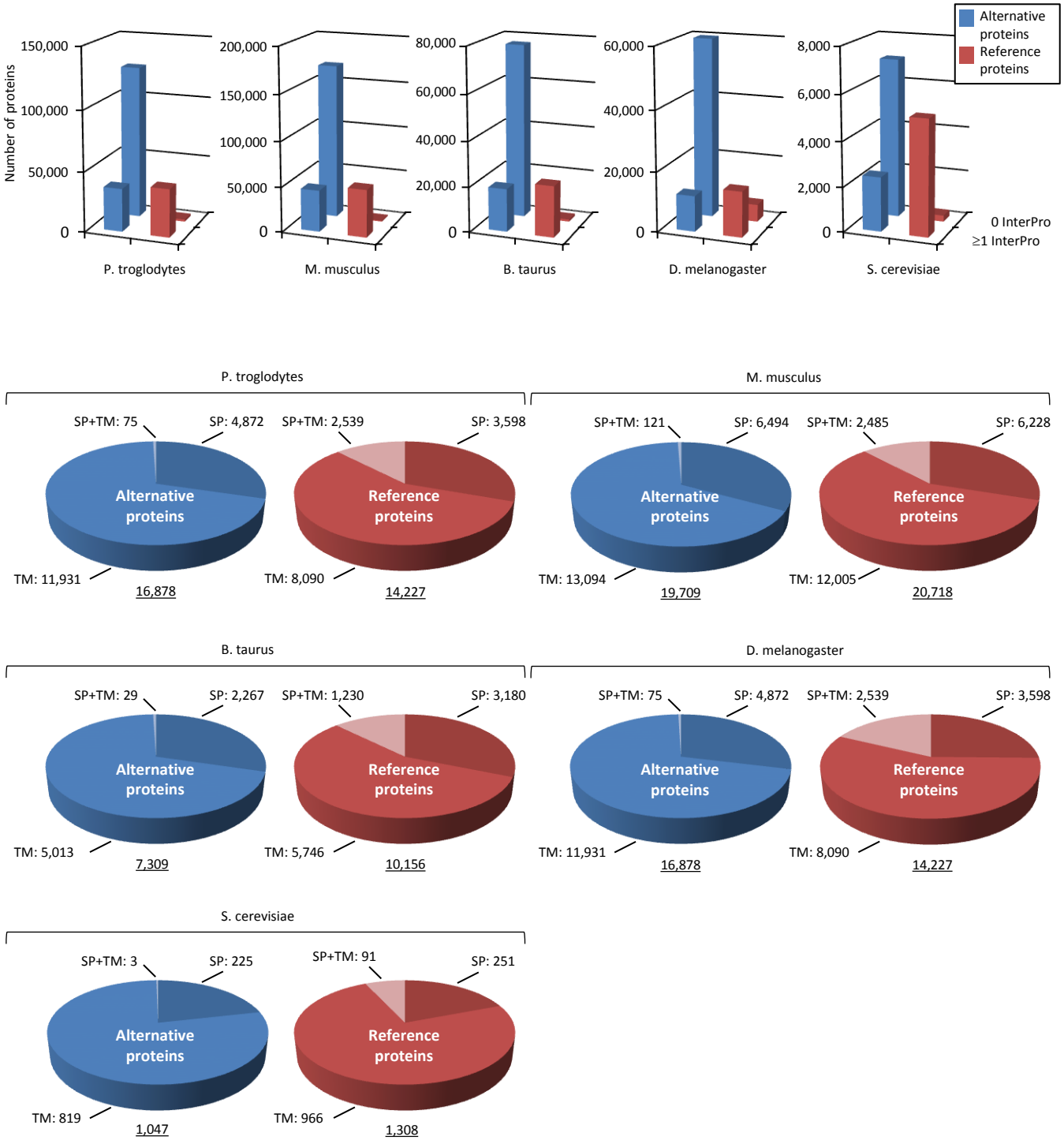


Figure 9-figure supplement 1: Alternative proteome sequence analysis and classification in *P. troglodytes, M. musculus, B. Taurus, D. melanogaster* and *S. cerevisiae*.
For each organism, the number of InterPro signatures (top graphs) and proteins with transmembrane (TM), signal peptide (SP), or TM+SP features (bottom pie charts) is indicated for alternative and reference proteins.

Figure 10

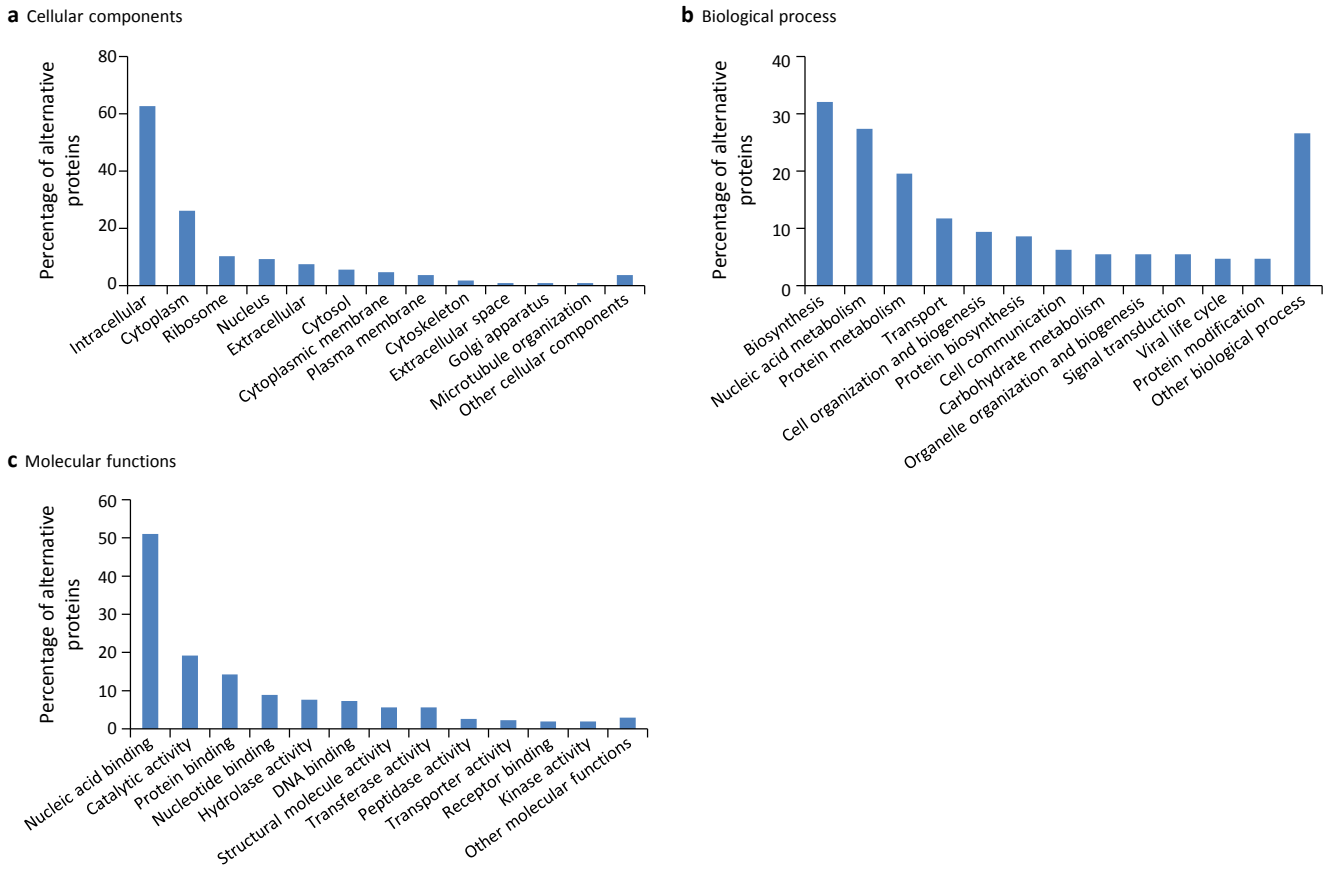


Figure 10. Gene ontology (GO) annotations for human alternative proteins.
GO terms assigned to InterPro entries are grouped into 13 categories for each of the three ontologies. (**a**) 34 GO terms were categorized into cellular component for 107 alternative proteins. (**b**) 64 GO terms were categorized 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 11

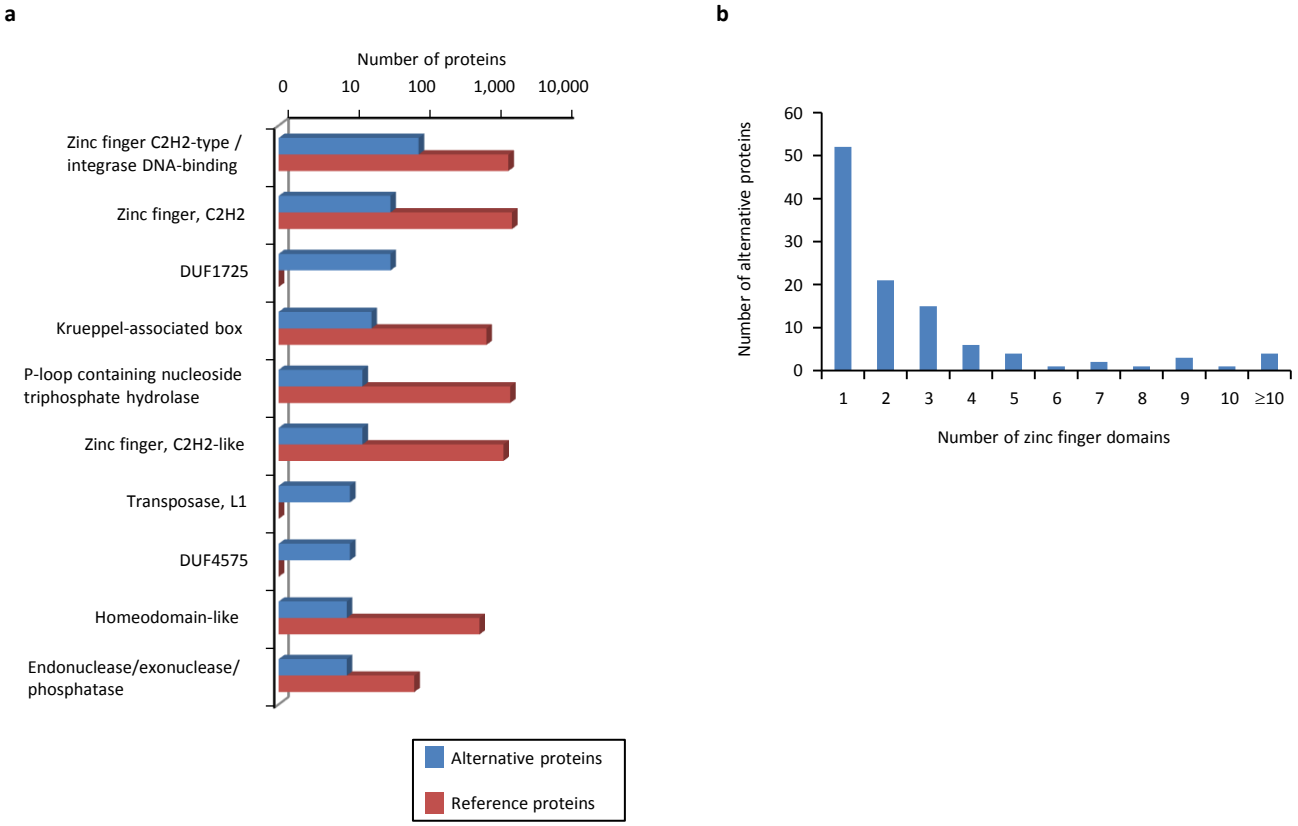


Figure 11. Main InterPro entries in human alternative proteins. (**a**) The top 10 InterPro families in the human alternative proteome. (**b**) A total of 110 alternative proteins have between 1 and 23 zinc finger domains.

Figure 12

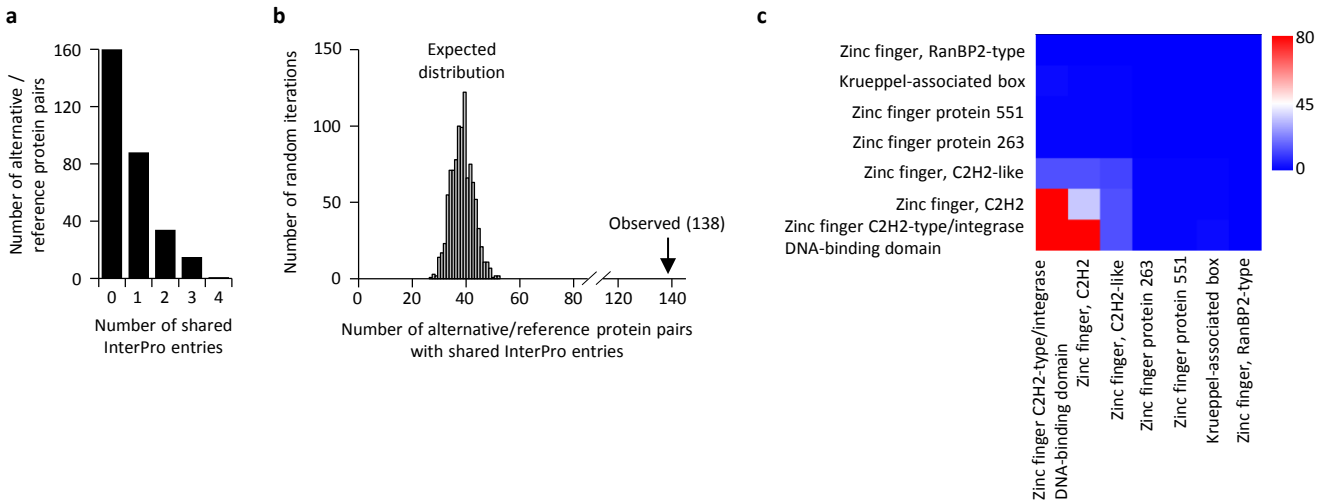


Figure 12. Reference and alternative proteins share functional domains.
(**a**) Distribution of the number of shared InterPro entries between alternative and reference proteins coded by the same transcripts. 138 pairs of alternative and reference proteins share between 1 and 4 protein domains (InterPro 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.

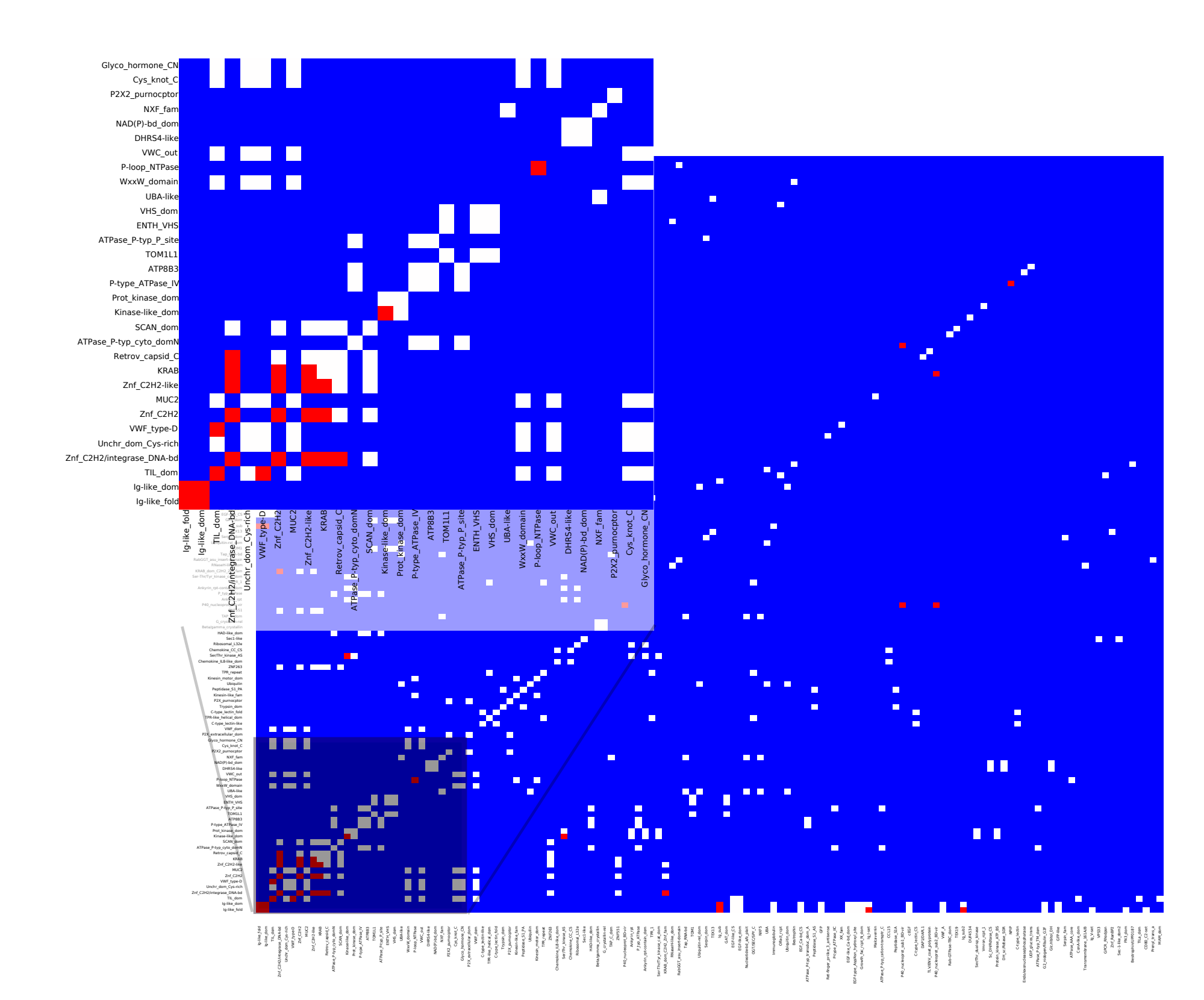

Glyco_hormone_CN
Cys_knot_C
P2X2_purnocptor
NXF_fam
NAD(P)-bd_dom
DHRS4-like
VWC_out
P-loop_NTPase
WxxW_domain
UBA-like
VHS_dom
ENTH_VHS
ATPase_P-typ_P_site
TOM1L1
ATP8B3
P-type_ATPase_IV
Prot_kinase_dom
Kinase-like_dom
SCAN_dom
ATPase_P-typ_cyto_domN
Retrov_capsid_C
KRAB
Znf_C2H2-like
MUC2
Znf_C2H2
VWF_type-D
Unchr_dom_Cys-rich
Znf_C2H2/integrase_DNA-bd
TIL_dom
Ig-like_dom
Ig-like_fold

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.

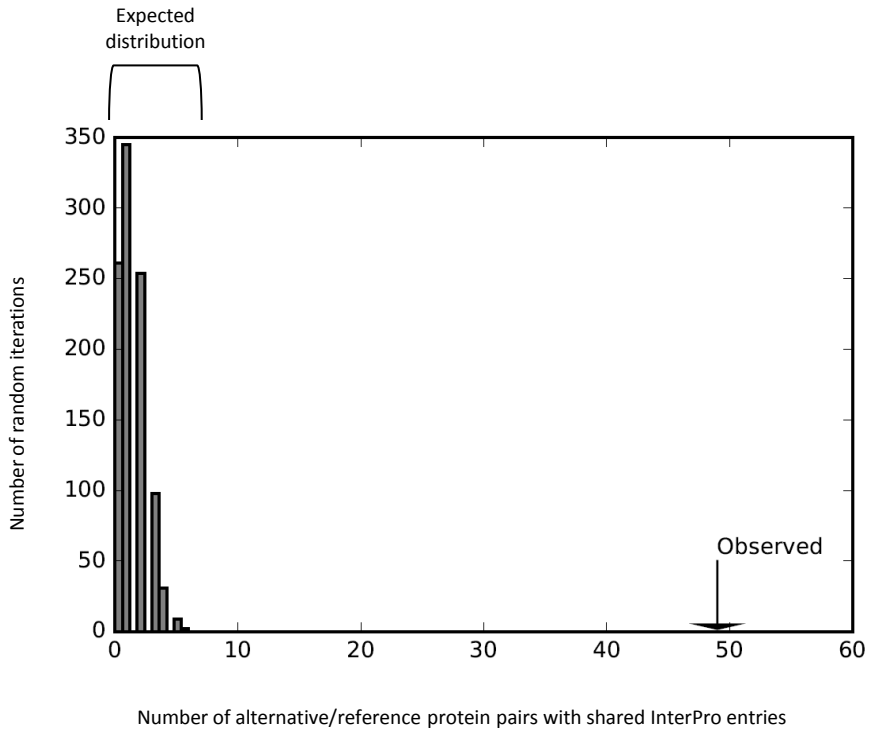


Figure 12-figure supplement 2: Reference and alternative proteins share functional domains.

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.

Figure 13

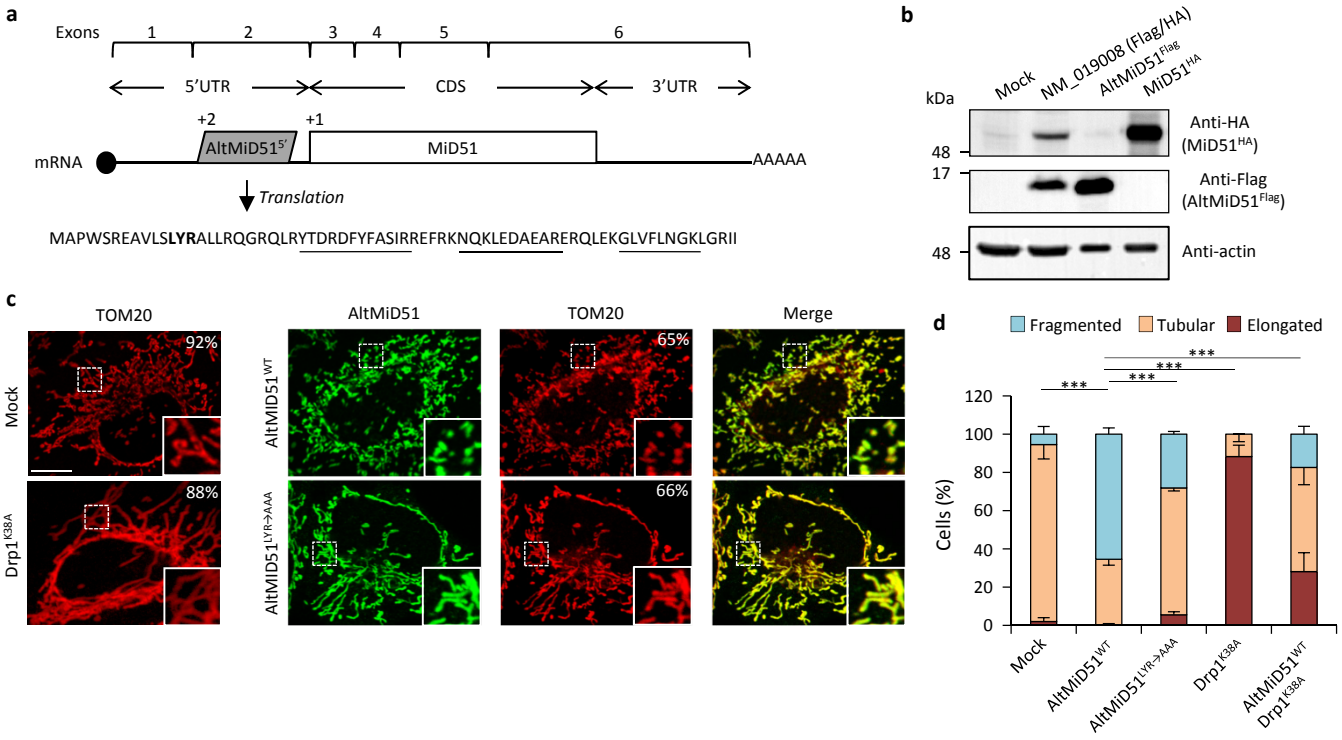
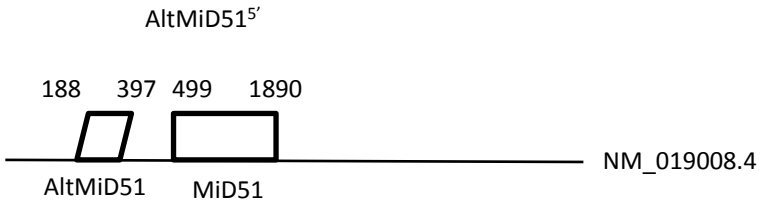
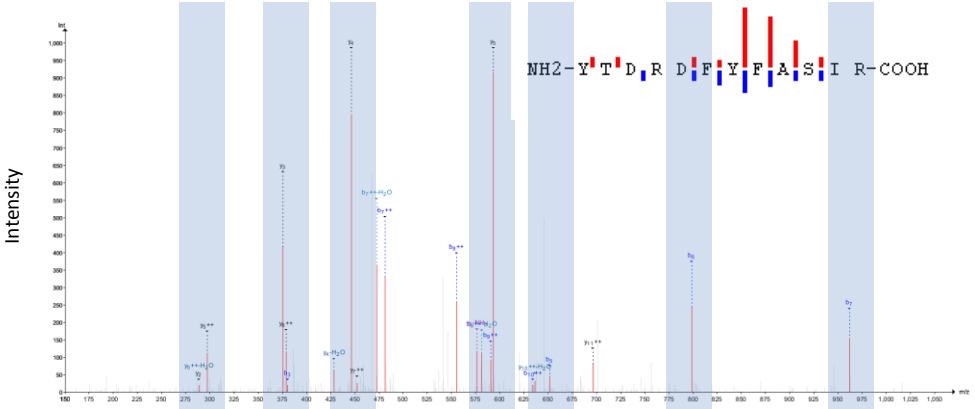


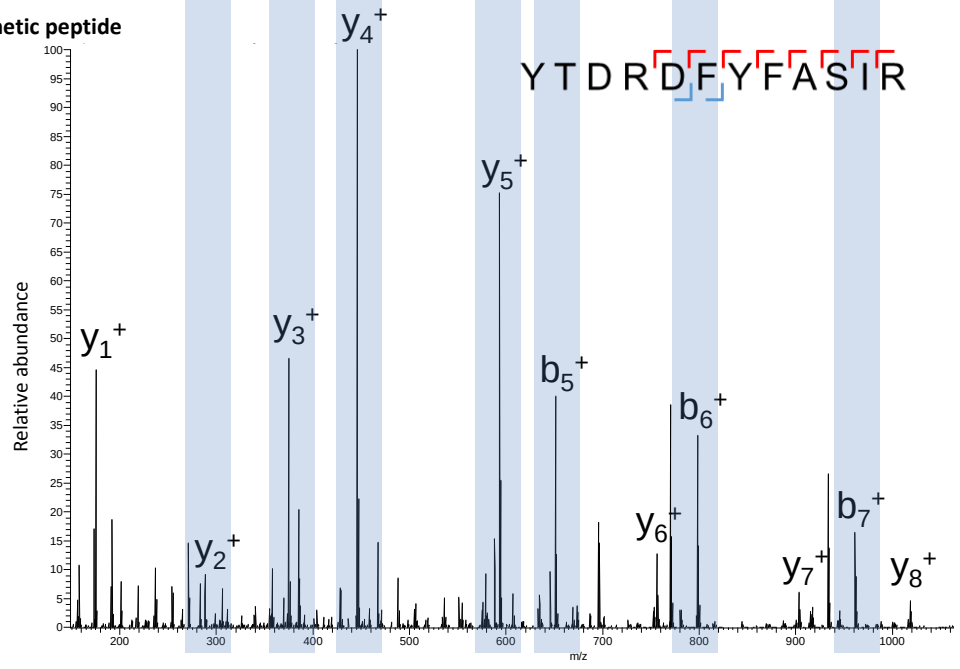
Figure 13. AltMiD51$^{5'}$ expression induces mitochondrial fission.
(**a**) AltMiD51$^{5'}$ coding sequence is located in exon 2 or the *MiD51/MIEF1/SMCR7L* gene and in the 5'UTR of the canonical mRNA (RefSeq NM_019008). +2 and +1 indicate reading frames. AltMiD51 amino acid sequence is shown with the 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 μm. (**d**) Bar graphs show mitochondrial morphologies in HeLa cells. Means of three independent experiments per condition are shown (100 cells for each independent experiment). ***p<0.0005 (Fisher's exact test) for the three morphologies between altMiD51(WT) and the other experimental conditions.



a. Experimental peptide



b. Synthetic peptide



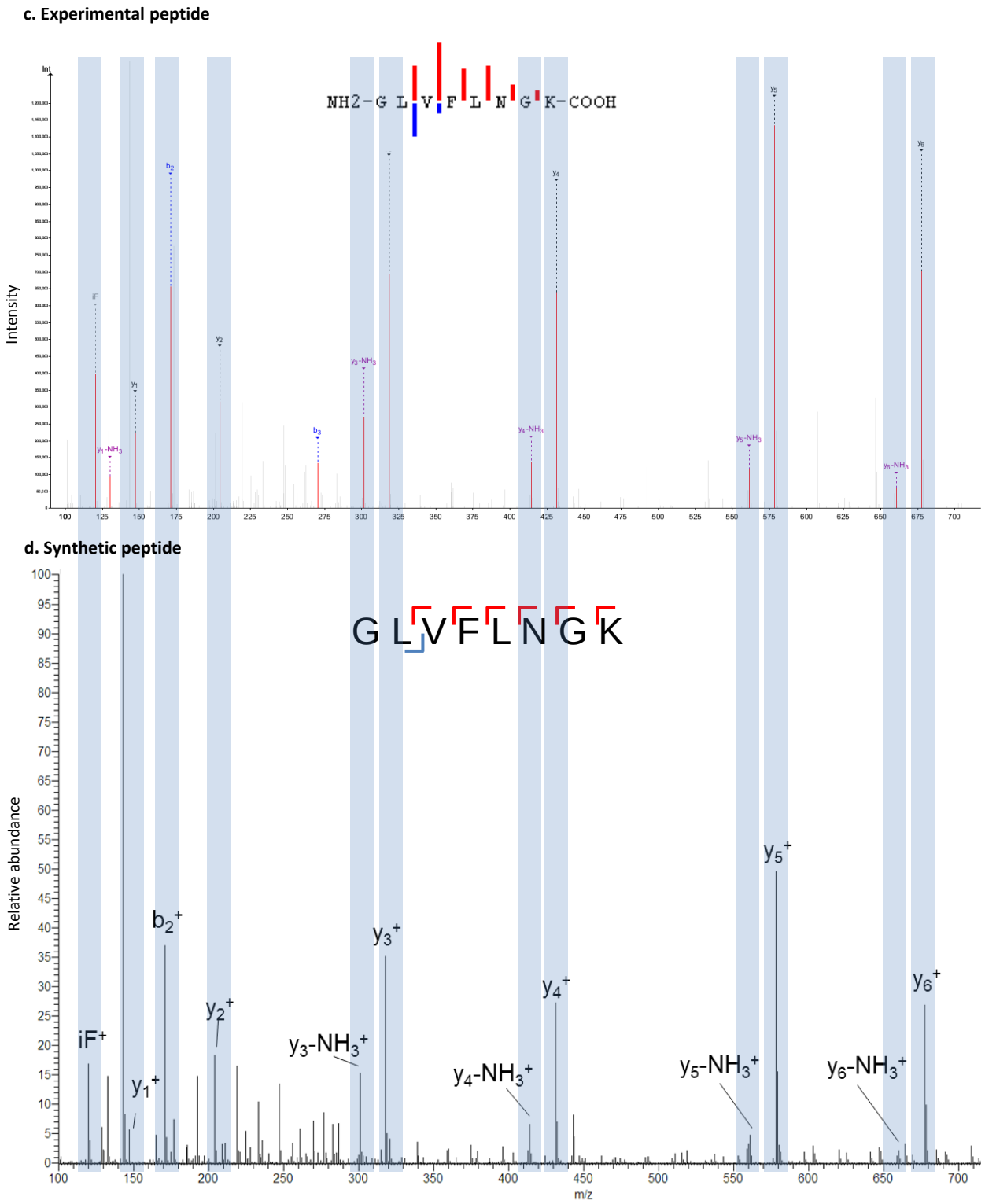


Figure 13-figure supplement 1: Spectra validation for altMiD51.
Example of validation for altMiD51 specific peptides YTDRDFYFASIR and GLVFLNGK. (**a,c**) Experimental MS/MS spectra (PeptideShaker graphic interface output). (**b,d**) MS/MS spectra of the synthetic peptides.
Matching peaks are shown with 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.

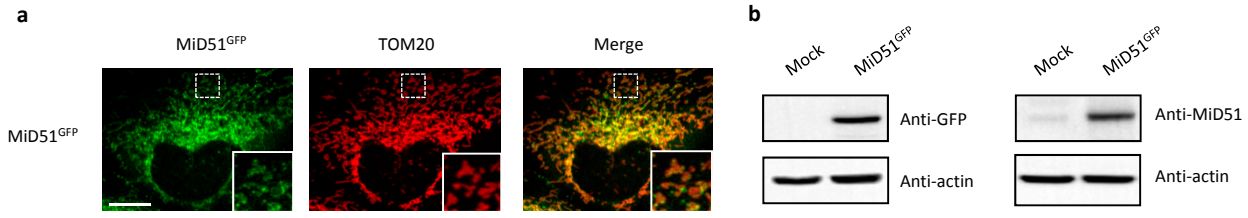


Figure 13-figure supplement 2: MiD51 expression results in mitochondrial fission. (**a**) Confocal microscopy of HeLa cells transfected with MiD51^{GFP} immunostained with anti-TOM20 (red channel) monoclonal antibodies. In each image, boxed areas are shown at higher magnification in the bottom right 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.

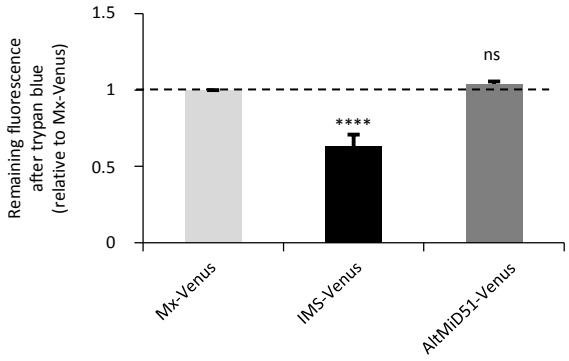


Figure 13-figure supplement 3: AltMiD51 is localized in the mitochondrial matrix. Trypan blue quenching experiment performed on HeLa cells stably expressing the indicated constructs. The fluorescence remaining after quenching by trypan blue is shown relative to Matrix-Venus (Mx-Venus) indicated by the 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.

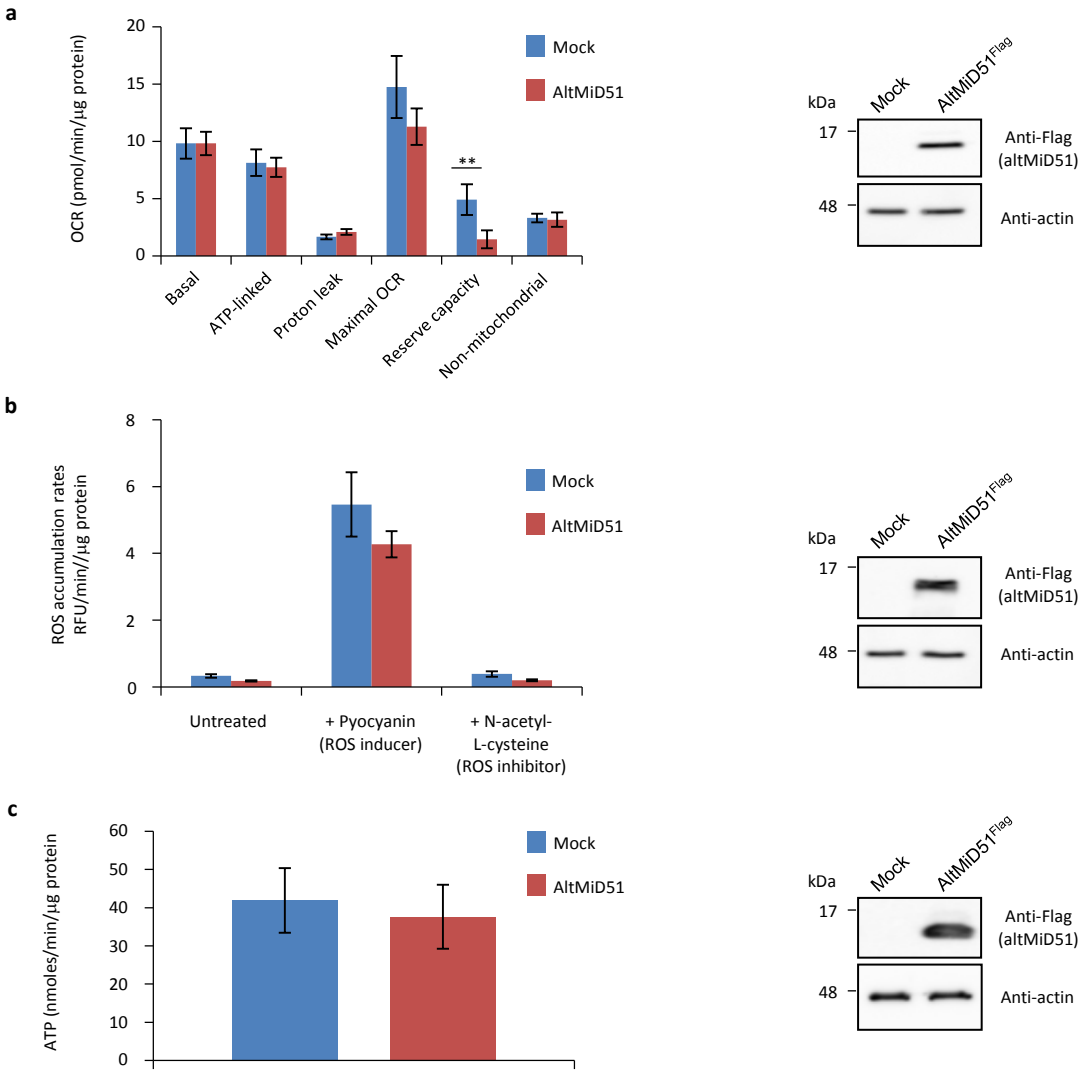


Figure 13-figure supplement 4: Mitochondrial function parameters.

(**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 three independent experiments (8 technical replicates each). 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).

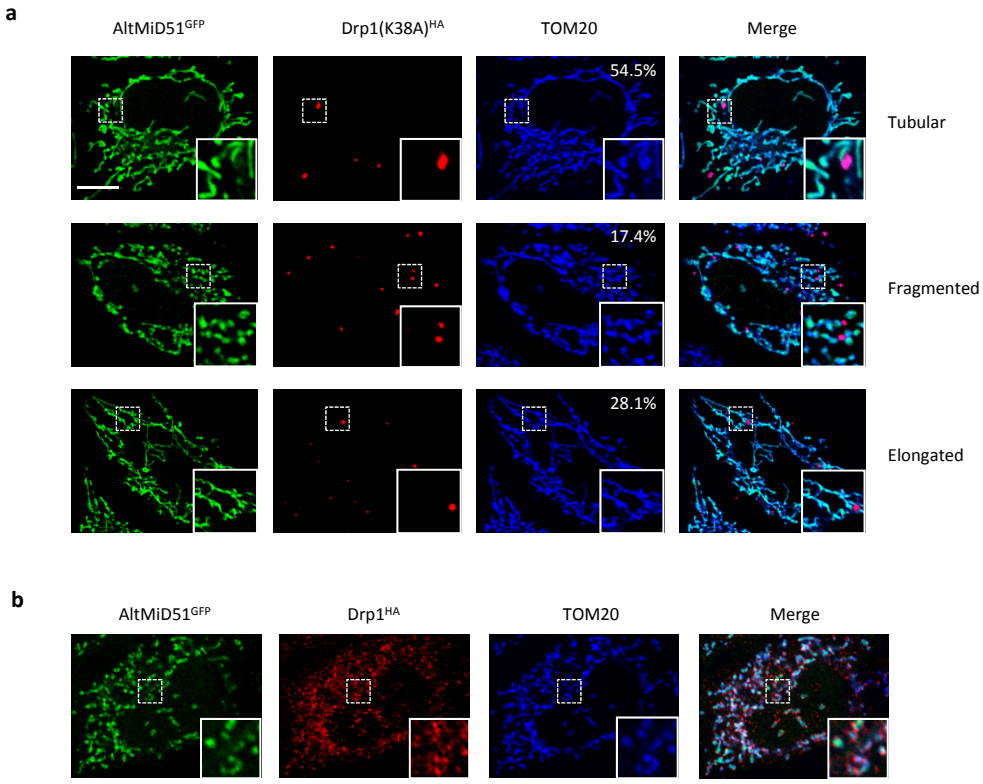


Figure 13-figure supplement 5: Representative confocal images of cells co-expressing altMiD51^{GFP} and Drp1(K38A)^{HA}.

(**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.

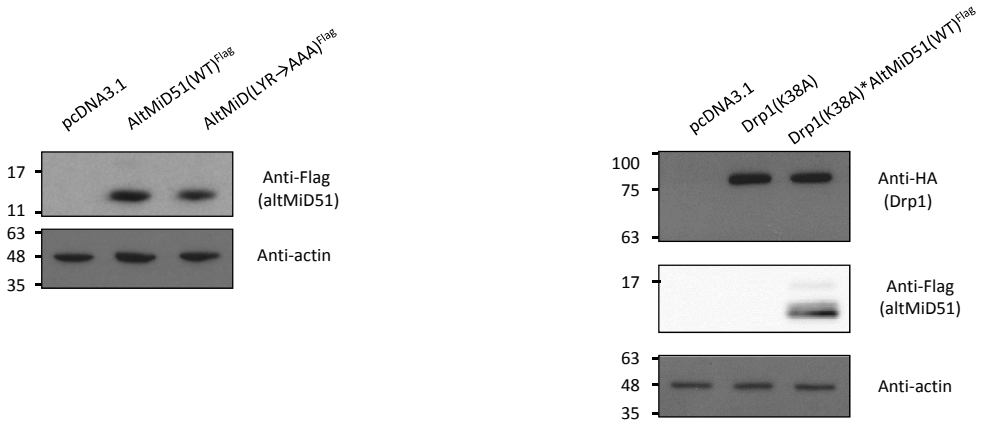


Figure 13-figure supplement 6: Protein immunoblot showing the expression of different constructs in HeLa cells.

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). Representative experiment of three independent biological replicates.

Figure 14

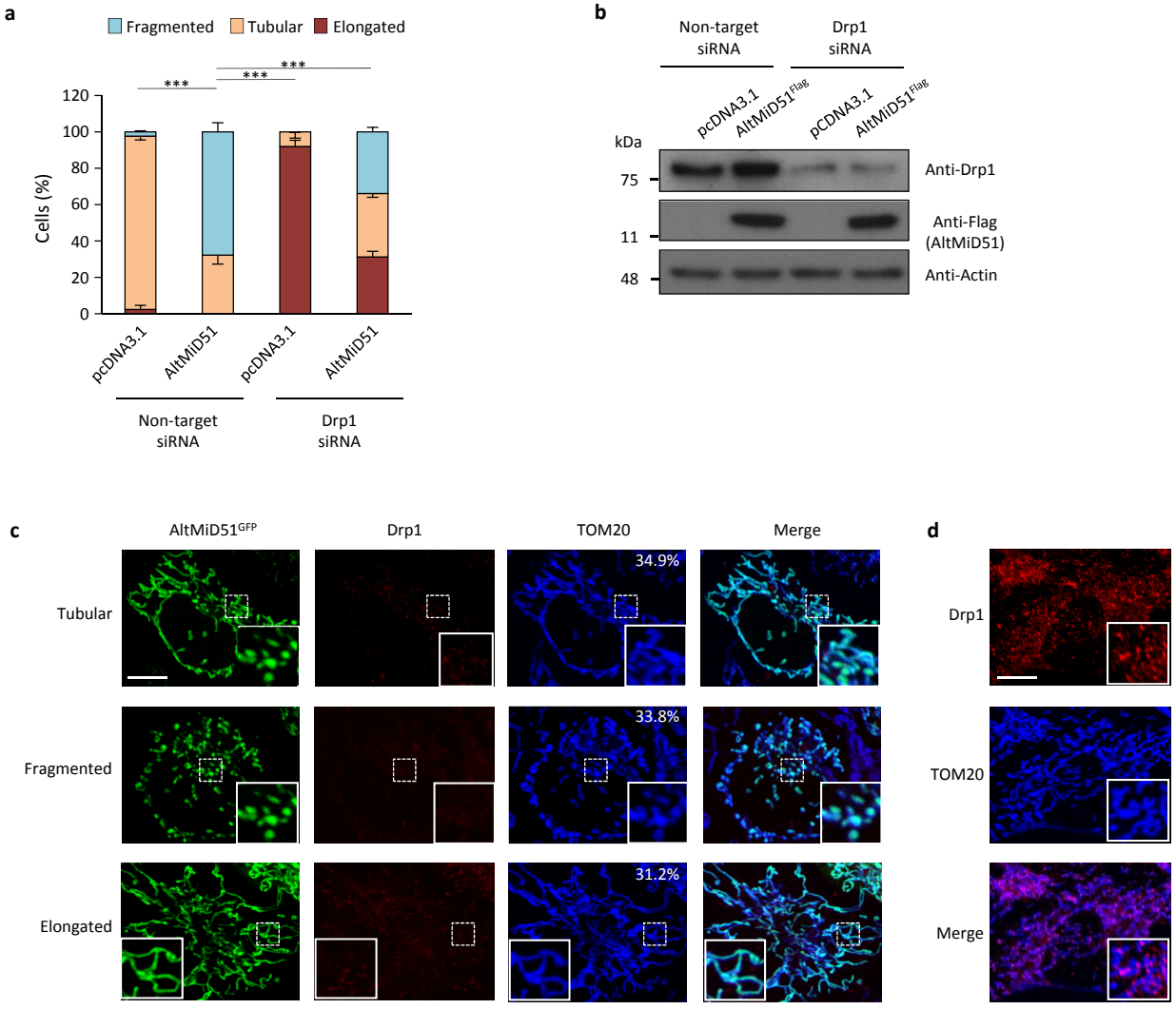


Figure 14. AltMiD51-induced mitochondrial fragmentation is dependent on Drp1. (**a**) Bar graphs show mitochondrial morphologies in HeLa cells treated with non-target or Drp1 siRNAs. Cells were mock-transfected (pcDNA3.1) or transfected with altMiD51^{Flag}. Means of three independent experiments per condition are shown (100 cells for each independent experiment). ***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. (**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 15

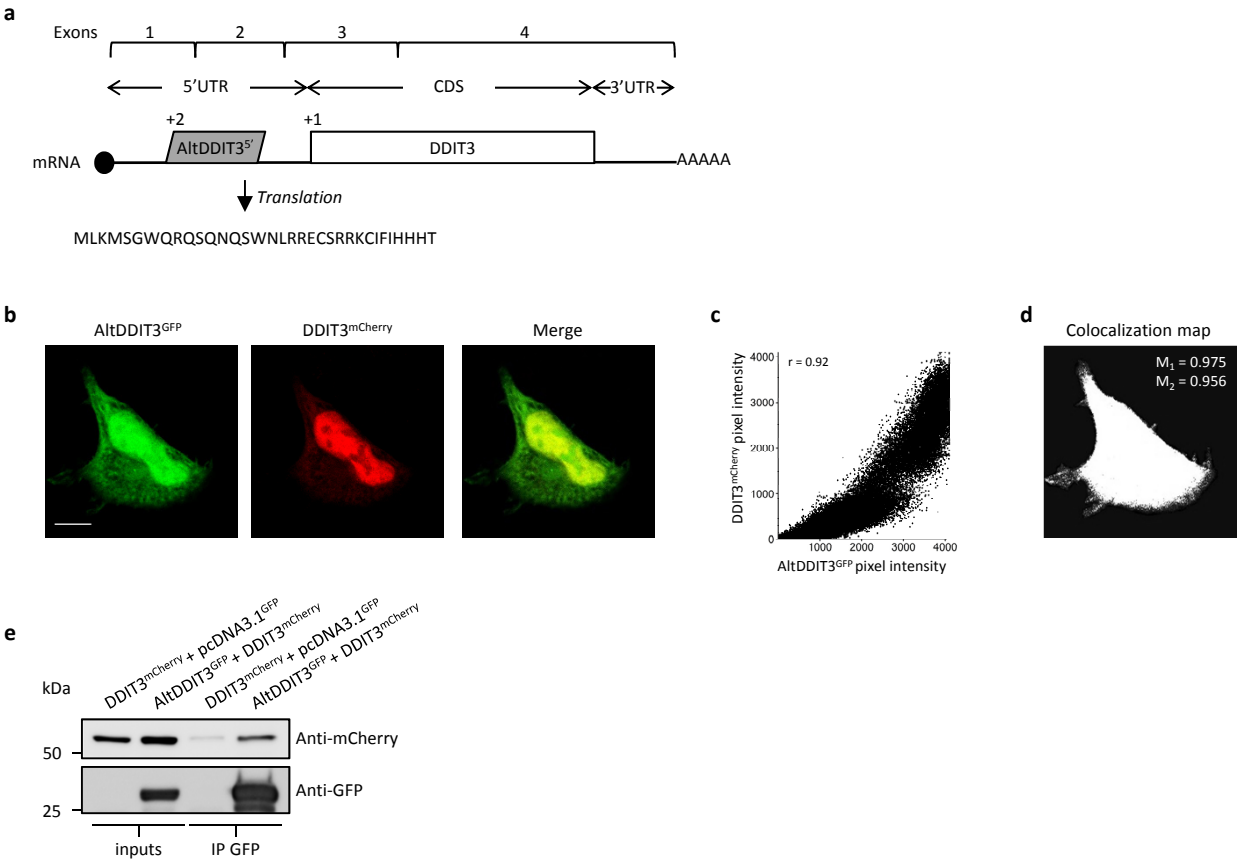


Figure 15. AltDDIT3$^{5'}$ co-localizes and interacts with DDIT3.
(**a**) AltDDIT3$^{5'}$ coding sequence is located in exons 1 and 2 or the *DDIT3/CHOP/GADD153* gene and in the 5'UTR of the canonical mRNA (RefSeq NM_004083.5). +2 and +1 indicate reading frames. AltDDIT3 amino acid sequence is also shown. (**b**) Confocal microscopy analyses of HeLa cells co-transfected with altDDIT3^{GFP} (green channel) and DDIT3^{mCherry} (red channel). Scale bar, 10 µm. (**c, d**) Colocalization analysis of the images shown in (**b**) performed using the JACoP plugin (Just Another Co-localization Plugin) implemented in Image J software (two independent biological replicates). (**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 (**b**) 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 (two independent experiments).

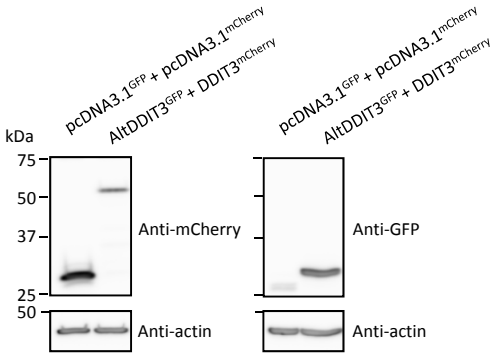


Figure 15-figure supplement 1: Protein immunoblot showing the expression of different constructs in HeLa cells.

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). Representative experiment of two independent biological replicates.

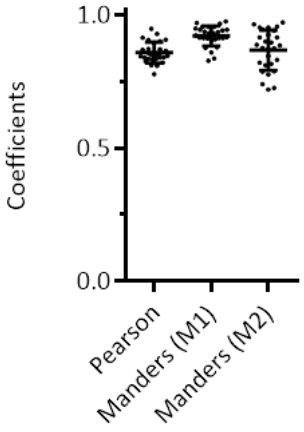


Figure 15–figure supplement 2 : Colocalization of altDDIT3 with DDIT3.
Scatter plots of Pearson's Correlation Coefficient and Manders' Correlation Coefficient after Costes' automatic threshold (*p*-value < 0.001, based on 1000 rounds of Costes' randomization colocalization analysis). M1 is the proportion of altDDIT3^{GFP} signal overlapping DDIT3^{mCherry} signal and M2 is the proportion of DDIT3^{mCherry} signal overlapping altDDIT3^{GFP}. Error bars represent the mean +/- SD of three independent experiments (28 cells).