Unusually efficient CUG initiation of an overlapping reading frame in *POLG* mRNA yields novel protein POLGARF

Gary Loughran[a,1], Alexander V. Zhdanov[a,1], Maria S. Mikhaylova[b,1], Fedor N. Rozov[c], Petr N. Datskevich[c], Sergey I. Kovalchuk[d], Marina V. Serebryakova[b], Stephen J. Kiniry[a], Audrey M. Michel[a], Patrick B. F. O'Connor[a], Dmitri B. Papkovsky[a], John F. Atkins[a], Pavel V. Baranov[a,d,2], Ivan N. Shatsky[b,2], and Dmitry E. Andreev[b,d,2]

[a]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland, T12 XF62; [b]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia, 119992; [c]Biological Faculty, Lomonosov Moscow State University, Moscow, Russia, 119234; and [d]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia, 117997

Edited by Eugene V. Koonin, National Institutes of Health, Bethesda, MD, and approved August 13, 2020 (received for review January 24, 2020)

While near-cognate codons are frequently used for translation initiation in eukaryotes, their efficiencies are usually low (<10% compared to an AUG in optimal context). Here, we describe a rare case of highly efficient near-cognate initiation. A CUG triplet located in the 5′ leader of *POLG* messenger RNA (mRNA) initiates almost as efficiently (∼60 to 70%) as an AUG in optimal context. This CUG directs translation of a conserved 260-triplet-long overlapping open reading frame (ORF), which we call *POLGARF* (*POLG* Alternative Reading Frame). Translation of a short upstream ORF 5′ of this CUG governs the ratio between POLG (the catalytic subunit of mitochondrial DNA polymerase) and POLGARF synthesized from a single *POLG* mRNA. Functional investigation of POLGARF suggests a role in extracellular signaling. While unprocessed POLGARF localizes to the nucleoli together with its interacting partner C1QBP, serum stimulation results in rapid cleavage and secretion of a POLGARF C-terminal fragment. Phylogenetic analysis shows that *POLGARF* evolved ∼160 million y ago due to a mammalian-wide interspersed repeat (MIR) transposition into the 5′ leader sequence of the mammalian *POLG* gene, which became fixed in placental mammals. This discovery of *POLGARF* unveils a previously undescribed mechanism of de novo protein-coding gene evolution.

uORF | 5′ leader | non-AUG initiation | start codon selection | dual coding gene

The process of translation can be described in four steps: initiation, elongation, termination, and ribosome recycling. It is believed that protein synthesis is mostly regulated at the level of initiation. In eukaryotes, the scanning model for translation initiation postulates that the small ribosomal subunit, in complex with initiation factors and Met-initiator transfer RNA ($tRNA_i$), enters at the 5′ end of messenger RNA (mRNA) and then scans toward the 3′ end (1, 2). Base-pairing interactions between the anticodon of the Met-$tRNA_i$ and an AUG codon in the mRNA halt ribosome scanning and set the reading frame for subsequent elongation steps (3). Notably, due to mRNA mispairing with the anticodon of Met-$tRNA_i$, initiation can also occur at most triplets that differ from AUG by a single nucleotide (near-cognate), albeit with much lower efficiency (4). However, some examples of highly efficient near-cognate initiation have been reported (5, 6).

Initiation efficiency on any translation initiation site (TIS) critically depends on its surrounding nucleotide context. In pioneering work, Kozak proposed that the context comprising six nucleotides (nt) before and one nt immediately following a potential initiation codon has significant influence on its recognition as a TIS (7). In agreement with Kozak, recent high-throughput analysis of all possible initiation contexts revealed RYMRMVAUGGC as the optimal context in human and mouse cells and additionally revealed synergistic effects of neighboring nucleotides (8).

TISs in unfavorable context can be bypassed by the scanning ribosome in a process known as leaky scanning. Since many mammalian mRNA 5′ leaders possess AUG codons, as well as many potential near-cognate start codons, then leaky scanning must be widespread. However, the mere presence of a potential TIS in a 5′ leader doesn't necessarily guarantee initiation there. Until recently, it was difficult to estimate how frequently upstream TISs (uTISs) are recognized by scanning ribosomes in living cells. This can now be directly addressed since the emergence of the ribosome profiling technique (Riboseq), which allows monitoring of global translation at single nucleotide resolution (9). Riboseq revealed widespread translation in the 5′ leaders of mRNAs, especially in mammalian cells (6, 10–12).

What is the role of translation initiation in 5′ leaders? In some instances, it gives rise to N-terminal extensions (13–15) although, in most cases, they result in translation of short (some are simply AUG-stop) upstream open reading frames (uORFs). While it is believed that most uORFs suppress translation of their main

Significance

In this study, we describe a previously unknown mechanism of de novo protein-coding gene evolution. We show that the *POLG* gene, which encodes the catalytic subunit of mitochondrial DNA polymerase, is in fact a dual coding gene. Ribosome profiling, phylogenetic conservation, and reporter construct analyses all demonstrate that *POLG* mRNA possesses a conserved CUG codon which serves as a start of translation for an exceptionally long overlapping open reading frame (260 codons in human) present in all placental mammals. We called the protein encoded in this alternative reading frame POLGARF. We provide evidence that the evolution of *POLGARF* was incepted upon insertion of an MIR transposable element of the SINE family.

Author contributions: G.L., A.V.Z., M.S.M., P.V.B., and D.E.A. designed research; G.L., A.V.Z., M.S.M., F.N.R., P.N.D., S.I.K., M.V.S., S.J.K., A.M.M., P.B.F.O., P.V.B., and D.E.A. performed research; D.E.A. contributed new reagents/analytic tools; G.L., A.V.Z., M.S.M., S.I.K., D.B.P., J.F.A., P.V.B., I.N.S., and D.E.A. analyzed data; and G.L., A.V.Z., M.S.M., D.B.P., J.F.A., P.V.B., I.N.S., and D.E.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]G.L., A.V.Z., and M.S.M. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: p.baranov@ucc.ie, ivanshatskyster@gmail.com, or cycloheximide@yandex.ru.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2001433117/-/DCSupplemental.

First published September 21, 2020.

protein-coding sequence (16–18), a number of uORFs are involved in more specialized regulation of translation, ranging from selective stress responses to eIF2 phosphorylation (19–22) to metabolite sensing (23–26).

Our attention was drawn to one such uORF within an mRNA encoding a catalytic subunit of mammalian mitochondrial DNA polymerase (POLG). POLG is a hotspot for more than 200 known mutations in humans that cause mitochondria-associated diseases: such as, progressive external ophthalmoplegia with mitochondrial DNA deletions, autosomal dominant 1 (PEOA1); sensory ataxic neuropathy, dysarthria, ophthalmoparesis (SANDO); Alpers–Huttenlocher syndrome (AHS); and mitochondrial neurogastrointestinal encephalopathy (MNGIE) (https://tools.niehs.nih.gov/polg/). Disease development is believed to result from a gradual depletion of mitochondrial DNA (mtDNA) due to polymerase dysfunction(s). Transgenic mice with a POLG mutation that causes proofreading deficiency (Polg mutator mouse) develop an mtDNA mutator phenotype characterized by low mtDNA copy number, decreased life span, and premature aging (27).

There is a single AUG within the *POLG* 5′ leader which is expected to initiate translation of a conserved 23-codon uORF. Here, we show that, contrary to expectations, removal of the upstream AUG suppresses translation of the *POLG* coding sequence. Exploring the unusual effect of the uORF mutation revealed highly efficient CUG initiation of a 260-codon-long alternative reading frame (−1) overlapping the *POLG* main ORF. Thus, the *POLG* mRNA turns out to be a dual coding messenger.

Results

A CUG Triplet Located Upstream of the *POLG* Coding Sequence Initiates Translation of a Long Overlapping Reading Frame. The 5′ leader of the *POLG* mRNA contains a 23-codon conserved AUG-initiated uORF. To determine whether translation of this uORF affects the synthesis of POLG, we fused the whole 5′ leader of *POLG* (plus 33 nt downstream of the main ORF start codon) to a Firefly luciferase (Fluc) reporter and explored the effect of preventing uORF translation on reporter activity (Fig. 1*A*). In general, uORF translation represses translation of the main ORF by decreasing the number of scanning 43S preinitiation complexes that reach the main ORF start codon (18); however, here, translation of the uORF enhances main ORF translation (Fig. 1*A*). In a search for potential explanations, we examined *POLG* in publicly available ribosome profiling data (28, 29). Within the *POLG* mRNA, the phase of triplet periodicity of ribosome footprints supports translation of an alternative reading frame (−1 frame) that overlaps the *POLG* coding sequence (CDS) (Fig. 1*B*). This −1 frame footprint density is higher than the density of footprints aligning to the *POLG* CDS and decreases abruptly at the first −1 frame stop codon that is located in exon 3 of *POLG* (Fig. 1*C*). Notably, this −1 frame stop codon is conserved across placental mammals (Fig. 1 *C*, *Middle*), and the pattern of synonymous substitutions in the *POLG* CDS before the −1 frame stop codon suggests dual coding in this region.

Since there are no AUG triplets that could initiate translation of the −1 frame ORF, we searched for conserved potential near-cognate initiation codons and, in exon 2, identified a CUG triplet located 52 nt upstream of the *POLG* CDS start codon (Fig. 2*A*). To test whether this CUG triplet can initiate translation in the alternative reading frame, we fused the 5′ leader of *POLG* to Fluc in the −1 frame (Fig. 2*B*). We observed robust −1 frame translation that almost doubled when translation of the uORF was abolished (Fig. 2*B*, constructs 1 and 2). Thus, translation of the uORF decreases −1 frame initiation and increases 0 frame (POLG) initiation (Fig. 1*A*), which suggests preferential translation reinitiation on the more distal initiation codon (0 frame) after uORF translation (Figs. 1*A* and 2*B*). Replacement of the predicted CUG start codon with a noninitiating CUA completely abolished −1 frame translation, strongly suggesting that this CUG triplet is the only −1 frame initiation codon (Fig. 2*B*, construct 3). We termed this long ORF (260 codons in humans), which extensively overlaps with the annotated *POLG* reading frame and starts at CUG, as *POLG Alternative Reading Frame* (*POLGARF*).

The *POLG* CUG Acts as a Highly Efficient Initiation Codon. Notably, in reporter constructs, initiation at the *POLG* CUG is ∼60% as efficient as an AUG in the same position (Fig. 2*B*, constructs 1 and 4), which is markedly higher than most of the values reported for CUG initiation (5 to 10% efficiency, also observed in Fig. 2*B*, constructs 6 and 8) (30). Several previous studies reported that stable RNA secondary structures positioned ∼12 to 15 nt 3′ of a poor context start codon can stimulate initiation (31, 32). RNAfold (33) predicts a 35-nt RNA stem loop (ΔG = −20.5 kcal/mol) that starts 13 nt 3′ of the *POLG* CUG initiation codon (*SI Appendix*, Fig. S1*A*). We generated a series of −1 frame reporters to determine sequences important for *POLG* CUG initiation efficiency (*SI Appendix*, Fig. S1 *B–D*) and found that almost the entire 5′ leader is dispensable for efficient −1 frame initiation (*SI Appendix*, Fig. S1*B*). *POLG* −1 frame reporters with only 57 nt 3′, that still retain sequences predicted to be important for maintaining the stem loop, initiate as efficiently as reporters with all 3′ nucleotides (*SI Appendix*, Fig. S1*C*). We observed a decrease from >60% to just 20% CUG initiation efficiency when only the 5′ half of the predicted stem loop was included and no further decrease in efficiency when only a single 3′ codon is retained (*SI Appendix*, Fig. S1*C*). Mutations predicted to disrupt the stem loop reduce initiation efficiency, and this decrease is reversed in reporters with compensatory changes expected to restore the stem loop (*SI Appendix*, Fig. S1*C*). A series of systematic nested deletions from the 3′ end provide further support for a stimulatory role for the stem loop (*SI Appendix*, Fig. S1*D*). Surprisingly, we observed a striking increase in CUG initiation efficiency with reporters missing a significant part of the 3′ half of the predicted stem (compare +35 and +38 in *SI Appendix*, Fig. S1*D*). However, we subsequently noticed that a downstream vector-derived sequence can base pair with a *POLG* sequence to restore stem loop formation (*SI Appendix*, Fig. S1*E*). This confirms the importance of a downstream secondary structure for *POLGARF* translation.

Since the 3′ deletions did not reduce *POLG* CUG initiation below 20% and a CUG reporter in optimal Kozak context initiates at only 7% in our experiments (Fig. 2*B*), we concluded that the 3′ sequences (beyond +3) cannot be solely responsible for the highly efficient *POLG* CUG initiation. It has been reported that initiation at near-cognate codons may be more dependent on local context than initiation at AUG codons (34). Therefore, it seems highly probable that such efficient *POLG* CUG initiation is heavily reliant on its surrounding nucleotide context. Particularly important is a G at the +4 position (where the first nucleotide of the start codon is +1), which has been recently shown to be critical for near-cognate initiation efficiency (34). In the context of the *POLG* CUG, the −3A and +4G are evolutionarily conserved (Fig. 2*A*), and, as expected, when we exchanged −3A to G and +4G to A in both CUG- and AUG-initiating reporters, initiation at the AUG codon decreased by 30%, but CUG-initiated translation dropped by 85% (*SI Appendix*, Fig. S2*A*). This reinforces the idea that near-cognates are more sensitive to changes at the −3 and +4 positions. However, our CUG reporter in optimal Kozak context that only initiates at 7% efficiency also has −3G and +4A. A comparison between the local context (−6 to +6) of *POLG* CUG and the Kozak CUG reveals differences at positions −2,−1, +5, and +6. A role for positions +5 and +6 in CUG initiation has been previously reported (35, 36). Exchanging these positions in the *POLG* context with those from the Kozak CUG decreased *POLG* CUG

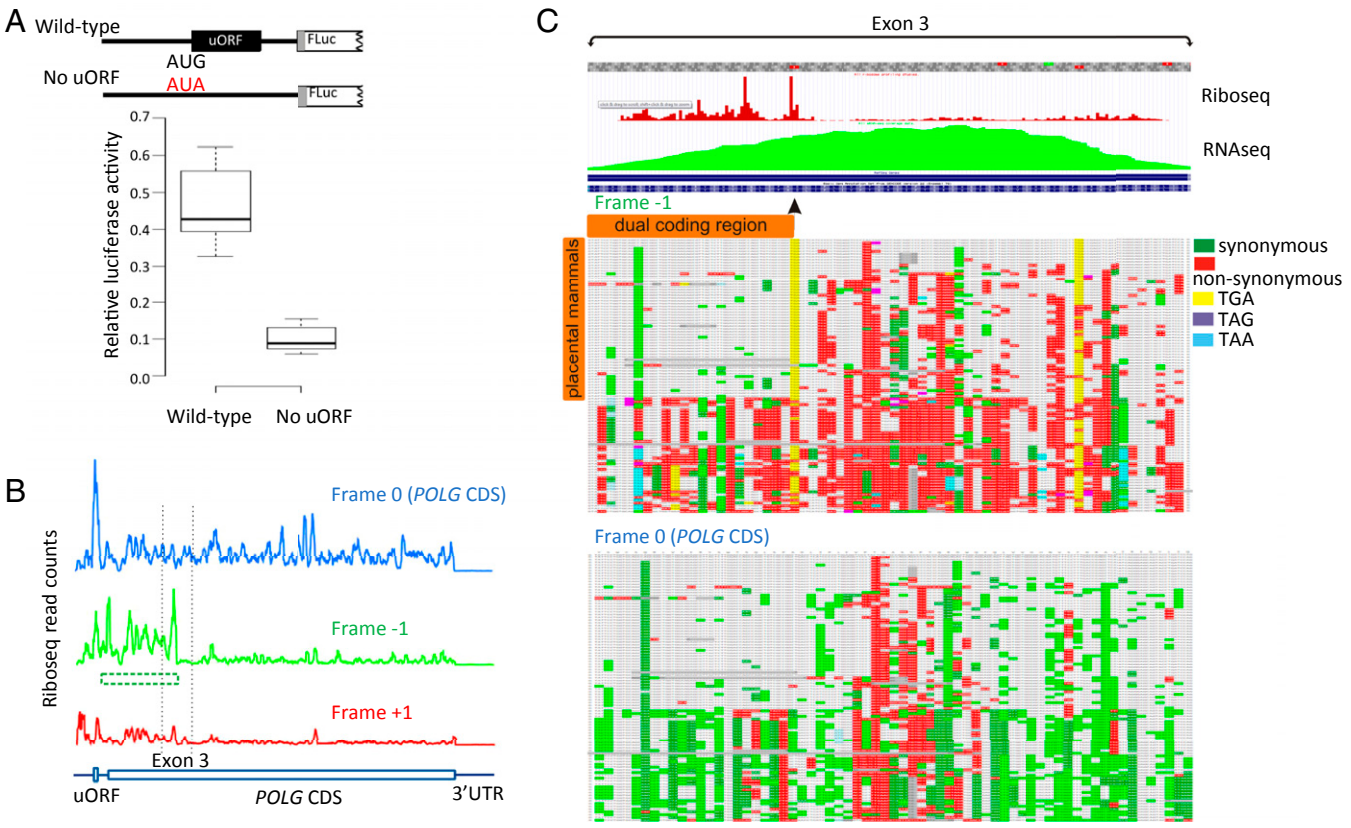

Fig. 1. Identification of an alternative translated ORF in *POLG*. (**A**) Schematic representation of reporter constructs bearing full-length *POLG* 5′ leaders (with and without the uORF AUG codon, mutated start codon is shown in red font), its AUG start codon plus 33 nt fused to firefly luciferase (in frame with the *POLG* AUG), and relative luciferase activities of corresponding constructs transfected into HEK293T cells ($n = 12$). (*B*) Riboseq aligned to each reading frame for the *POLG* mRNA generated in the Trips-Viz browser. The region in the −1 frame, which has high Riboseq density, is depicted by a green dotted box. Dotted vertical lines show exon 3 boundaries. (*C*) GWIPS-Viz tracks (29) for Riboseq (red) and RNA sequencing (RNAseq) (green) global aggregates for exon 3 of *POLG* (*Top*). *Middle* and *Bottom* represent CodAlignView alignment of 100 vertebrate genomes (hg38/100) for the −1 frame (*Middle*) and 0 frame (*Bottom*). Black arrowhead shows position of −1 frame stop codon.

initiation to 11% (*SI Appendix*, Fig. S2*B*). Both −2,−1 and +5,+6 nucleotide positions alone reduced CUG initiation although −2,−1 has much more of an impact (*SI Appendix*, Fig. S2*B*). These data support the idea that the optimal nucleotide context for near-cognate initiation is not the same as the optimal context for AUG initiation (34), and the finding that the *POLG* CUG initiation codon context is favorable for efficient near-cognate initiation. Furthermore, we tested several naturally occurring near-cognate contexts from *eIF4G2* GUG (37), *R3HCC1* CUG (13), and *TEAD1* AUU (38), which all have a high degree of similarity to the *POLG* CUG context (*SI Appendix*, Fig. S3*A*). All of these near-cognate codons initiated more efficiently than expected if they were in optimal Kozak context (*SI Appendix*, Fig. S3*B*). Replacement of the *eIF4G2* GUG codon with CUG increased initiation efficiency almost fourfold, verifying previous findings (6) and indicating that CUG is the preferred near-cognate start codon. Notably, *eIF4G2* CUG initiation is almost twofold more efficient than for *POLG* CUG without its stem loop.

Since all near-cognate initiation codons are expected to be suboptimal, it seemed likely that the *POLG* CUG initiation should be sensitive to the levels of initiation factors responsible for the stringency of suboptimal start codon selection: i.e., EIF1 and EIF5 or its antagonists BZW/5MP (6, 39–42). To test this, we coexpressed *POLG* reporters with an excess of either EIF1 or EIF5. To avoid the confounding effects of initiation factors' overexpression on reinitiation, *POLG* reporters were designed without the uORF. Surprisingly, we saw almost no effect of EIF1 or EIF5 overexpression on *POLG* CUG initiation. Thus, in its natural nucleotide context, the *POLG* CUG codon behaves as a canonical AUG codon in good Kozak context, which is also unresponsive to EIF1 and EIF5 overexpression (41, 43). Interestingly, reducing *POLG* CUG initiation by either deletion of the *POLG* 3′ sequences or placing the *POLG* CUG in a suboptimal context reverses its independence from EIF1 and EIF5 levels. (Fig. 2*C*). Thus, initiation efficiency rather than initiation codon identity appears to be the major determinant of sensitivity to EIF1/5 levels. This is in contrast to the effect of eIF5 and 5MP expression on GUG or CUG initiation from the *eIF4G2* context (6).

What is the role of the 23-codon uORF upstream of *POLG* CUG? It is likely that ribosomes access the *POLG* CUG either by leaky scanning past the uORF AUG and/or by reinitiation of ribosomes that have translated the uORF. To determine whether leaky scanning is important for *POLG* CUG initiation, we tested reporters in which reinitiation on CUG is prevented by extending the uORF from 23 to 71 codons. Compared to wild-type constructs, CUG initiation is reduced by >50%, suggesting that leaky scanning plays an important role in *POLG* CUG initiation (*SI Appendix*, Fig. S4). In accordance with this, reducing leaky scanning by inserting a second in-frame AUG within the extended uORF almost completely abolished CUG initiation. Confirmation that reinitiation plays an equally important role in CUG initiation is observed in constructs in which leaky scanning is prevented by insertion of an in-frame AUG (reinitiation is still possible). Here, we still observed CUG initiation that is approximately half of wild-type CUG initiation. Furthermore, reducing uORF length is thought to be more permissive for

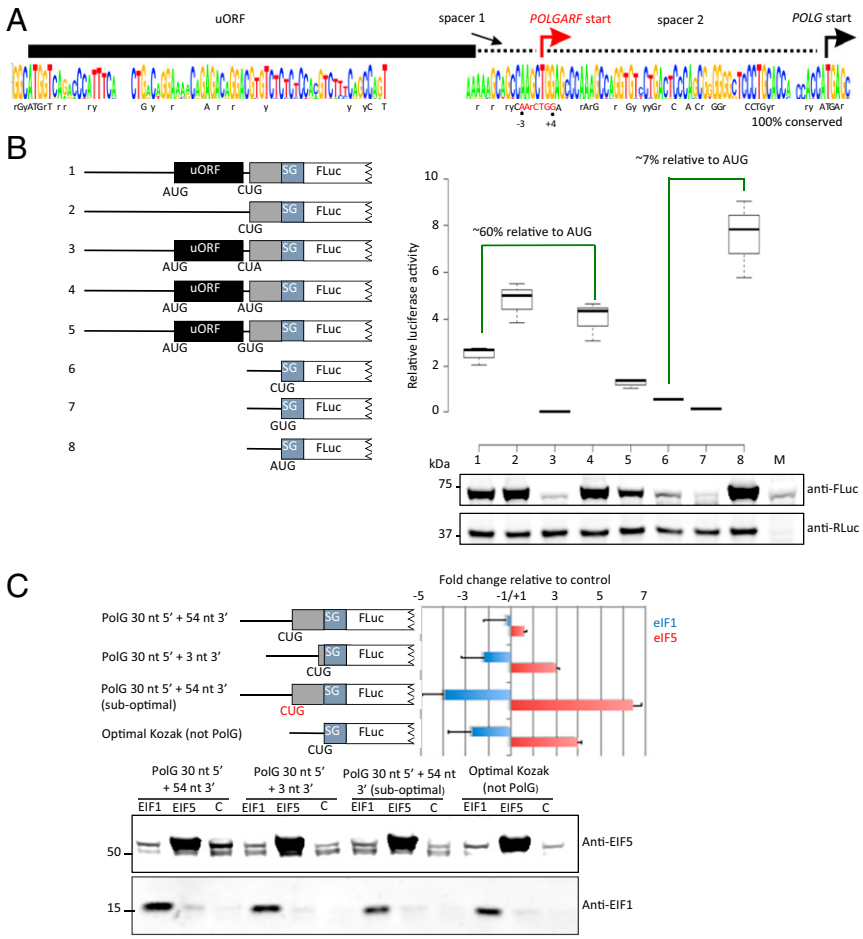

Fig. 2. *POLGARF* translation from a CUG codon. (*A*) Sequence logo (68) of multiple alignments of a *POLG* 5′ leader fragment generated for 85 placental mammals. Sequence below represents nucleotide positions that are universally conserved. The *POLGARF* CUG codon and surrounding nucleotides are highlighted in red. The r and y stand for purines and pyrimidines, respectively. (*B*) Schematic representation of reporter constructs transfected into HEK293T cells (*Left*) and relative luciferase activities (*Upper Right*, $n = 12$) and a representative anti-firefly luciferase and anti-Renilla luciferase (cotransfected with test constructs) Western blot (*Lower Right*). SG, porcine enterovirus StopGo; M, mock transfected. (*C*) Fold change in relative luciferase activities of selected constructs when cotransfected with plasmids expressing EIF1 or EIF5 relative to a control empty plasmid. Negative values indicate repression. The "sub-optimal" construct has a double substitution in the *POLG* CUG context (−3A/G and +4G/A). Shown are representative anti-EIF5 and anti-EIF1 Western blots (*Lower*) indicating overexpression.

reinitiation (44), and, consistent with this, we observed an ∼10% increase in CUG initiation by decreasing the uORF from 23 to 5 codons. In conclusion, both leaky scanning and reinitiation at the uORF appear to play equally important roles in *POLG* CUG initiation (*SI Appendix*, Fig. S4).

Being Originated through Mammalian-Wide Interspersed Repeat Transposition, POLGARF Is Conserved in Placental Mammals. To explore whether the *POLG* −1 frame ORF encodes a functional protein, we carried out phylogenetic analysis of 100 vertebrate genomes (45). The *POLG* −1 frame ORF as well as its CUG start codon are conserved in placental mammals, and PhyloCSF analysis (46) reveals strong purifying selection acting on the evolution of the POLGARF protein-coding sequence (Fig. 3*A*).

RepeatMasker (http://repeatmasker.org) identified a mammalian-wide interspersed repeat (MIR) within the *POLG* 5′ leader that overlaps with the short uORF and the CUG codon (Fig. 3*A*). Detailed analysis of vertebrate alignments suggests that the *POLG* −1 frame ORF originated by an MIR transposition before the adaptive radiation of placental mammals (Fig. 3*B* and *Discussion* below).

Detection of Endogenous POLGARF Protein. Alternative reading frame translation does not necessarily mean the existence of a stable protein product. To search for endogenous POLGARF peptides, we applied the post-acquisition targeted search (PATS) technique (47) to a BioPlex interactome dataset containing affinity purification-mass spectrometry results for ∼6,000 protein baits overexpressed in HEK293 cells (48). The PATS algorithm predicted POLGARF peptides in >90 protein baits, among which TRIP13, CAMK2D, NPM2, HAVCR2, CLEC3A, and CHCHD10 pull down datasets were predicted to have two or more POLGARF originated peptides (*SI Appendix*, Table S1).

Direct .raw data analysis for the latter baits indeed identified four POLGARF tryptic peptides; three of these are not found within the nr protein database by BLAST (49) and are unique for POLGARF. We confirmed the fidelity of peptide identification by comparing MS2 spectra from BioPlex .raw files with our liquid chromatography tandem mass spectrometry (LC-MS/MS) data produced with overexpressed POLGARF (*SI Appendix*, Fig. S5). The pattern of tryptic peptide coverage as well as the MS2 spectra were highly similar between the two datasets. Together these data unambiguously identify endogenous POLGARF protein in HEK293 cells from the BioPlex data.

Notably, the abundance of endogenous POLGARF in HEK293 cells seems to be very low as we were unable to detect

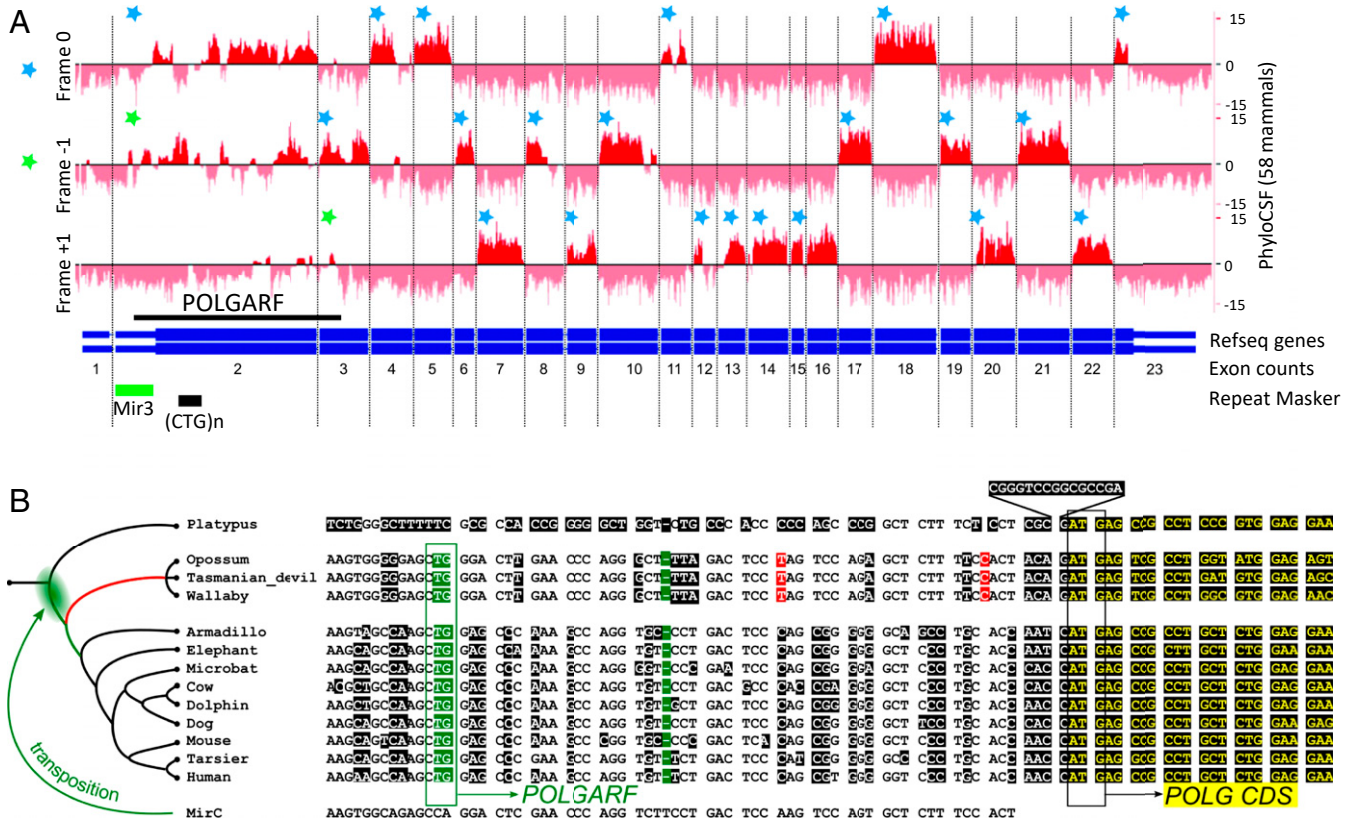

Fig. 3. Conservation and evolution of *POLGARF*. (*A*) PhyloCSF tracks for the *POLG* gene. *POLGARF* is indicated with a thick black line above RefSeq annotation. RepeatMasker tracks for *POLG* exons are below [$(CTG)_n$ corresponds to a poly-glutamine track in *POLG* for minus strand]. Reading frames for *POLG* and *POLGARF* ORFs are indicated with blue and green stars, respectively. (*B*) Possible origin of *POLGARF*. A genomic alignment is shown for representative mammalian species. The Dfam/Repbase MirC consensus is given below. Nucleotides that differ from the MirC sequence are highlighted: common among marsupials and placental mammals are in green; marsupial variants incompatible with *POLGARF* expression are in red; the rest are in black. *POLG* CDS sequence is in yellow. The timing of MirC sequence transposition is shown in the tree as a green cloud; the likely timing of mutations that led to the highlighted variants (in green and red) are indicated in the tree with the same colors. The initiation starts for POLG and POLGARF are indicated with rectangles.

POLGARF originated peptides from either in-house generated proteomics data without POLGARF overexpression or in deep proteome datasets from ref. 50 for HEK293, HeLa, colon, liver, HCT116, and prostate cells while, in the same datasets, several POLGARF originated peptides were detected in MCF7, SHSY, and A549 cells, as well as in Jurkat cells in the dataset from ref. 51 and in immune cells in the dataset from ref. 52. All peptide identifications are isolated and with very low intensities. Considering the high efficiency of CUG initiation from the *POLG* mRNA observed with Riboseq and reporter assays, it seems likely that endogenous POLGARF is either compartmentalized within subcellular locations that are insoluble under standard cell lysis protocols, secreted, or else unstable.

POLGARF Interacts with C1QBP and Redirects It to the Nucleoli. In order to shed light on POLGARF function, we searched for its interacting partners using a GST-POLGARF fusion protein overexpressed in Expi293F cells. Pull down assays detected a 32-kDa major protein identified as C1QBP, also known as P32 (Fig. 4*A*). The P32 homotrimer adopts a doughnut-shaped quaternary structure with asymmetric charge distribution on its surface. P32 is involved in a wide range of intracellular and extracellular activities. However, directed by a mitochondrial targeting sequence (MTS), it predominantly localizes in the mitochondrial matrix (53). In mitochondria, P32 is thought to control the translation of mitochondrially encoded proteins, either directly or by affecting mitochondrial ribosome biogenesis (54, 55). We confirmed the specificity of the POLGARF/P32 interaction using SNAP-tag pull down assays: SNAP-POLGARF and SNAP-P32 fusion proteins efficiently pulled down P32 and POLGARF, respectively (Fig. 4*B*). Next, we found that SNAP-POLGARF, when overexpressed in HEK293T cells, accumulates in nucleoli where it colocalizes with fibrillarin, one of the major nucleolar components (Fig. 4*C*), and does not colocalize with SC35, a nuclear speckle marker (*SI Appendix*, Fig. S6). Nucleolar localization of POLGARF is dependent on amino acids located within its N-terminal half (*SI Appendix*, Fig. S7). In nucleoli, POLGARF localizes to areas of active ribosomal RNA (rRNA) production, enriched with RNA polymerase I (Fig. 4*C*). SNAP-P32 alone was not observed in the nucleoli; however, when coexpressed with CLIP-POLGARF, SNAP-P32 showed clear nucleolar localization (Fig. 4*D*). To determine whether it is full-length, or mature P32 (without MTS) that accumulates in the nucleoli in a POLGARF-dependent manner, we carried out subcellular fractionation of cells overexpressing P32 with or without POLGARF (Fig. 4*E*). Mass spectrometry analysis of nucleolar P32 demonstrated that it retained the MTS (*SI Appendix*, Fig. S8). This suggests that its interaction with POLGARF prevents P32 maturation, redirects P32 from the mitochondria to the nucleoli, and thus may affect P32 functions.

A POLGARF C-Terminal Fragment, POLGARFin, Is Secreted from Cells upon Serum Stimulation. To investigate the kinetics of POLGARF accumulation, we fused POLGARF with HiBiT, an 11-amino acid peptide which can complement a truncated Nanoluciferase fragment (LgBiT) to regain full activity (56). After transfection

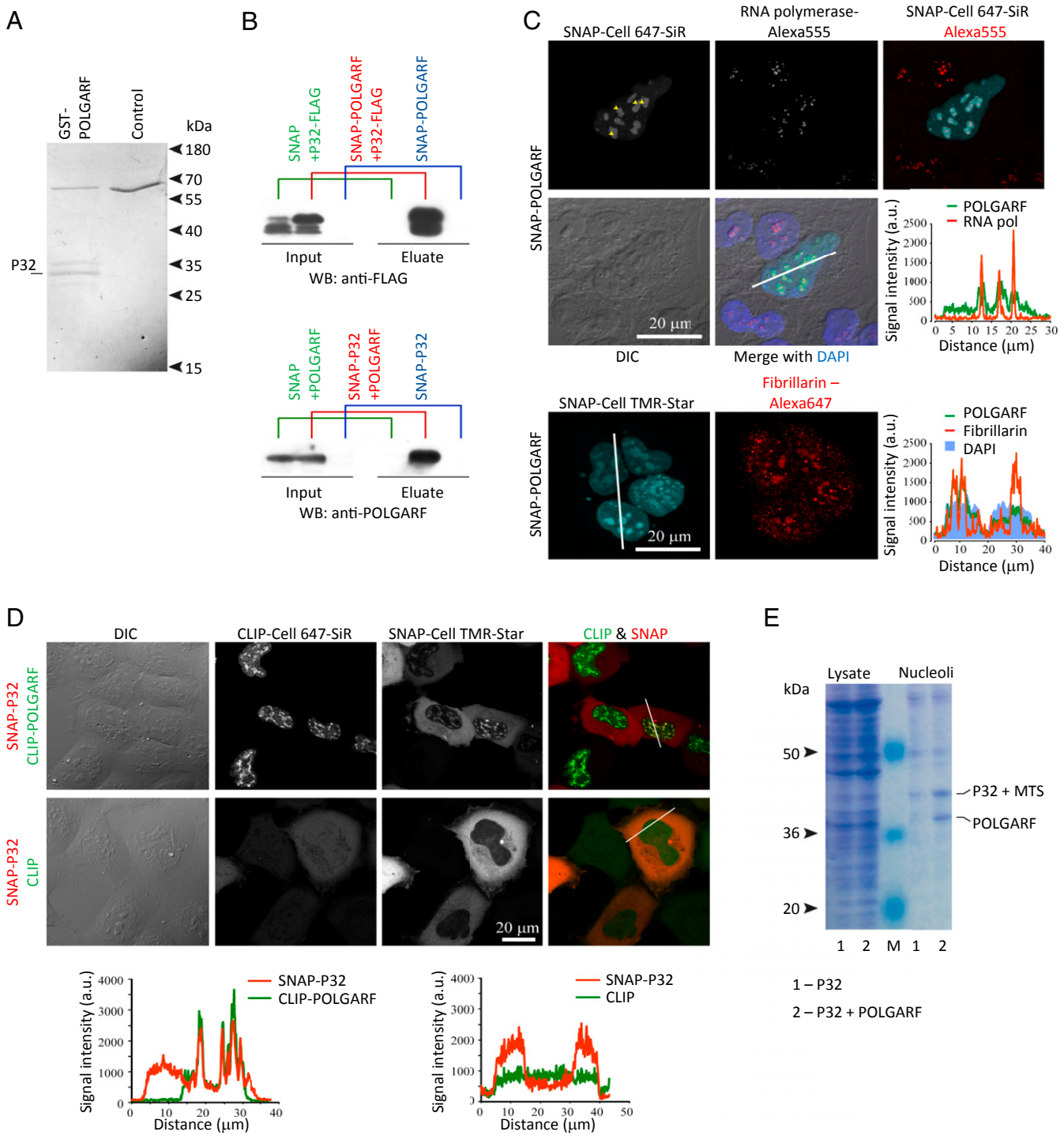

Fig. 4. POLGARF interaction, localization, and functional analysis. (*A*) Coomassie blue stained SDS/PAGE of a GST-POLGARF pull down assay. The major protein bands were excised and subject to mass spectrometry analysis. (*B*) SNAP pull down assay from cells transfected with indicated constructs. Anti-FLAG and anti-POLGARF antibodies (see also *SI Appendix*, Fig. S9) were used for Western blotting (WB) analysis. (*C* and *D*) Confocal imaging of cells transfected with the indicated constructs and stained with corresponding SNAP and CLIP dyes and antibodies, and differential interference contrast (DIC) images. Signal intensity is measured in arbitrary units (a.u.). Line profile analysis across representative cells (indicated by white lines) shows localization of fluorescent signals. In *C*, yellow arrowheads show localization of RNA polymerase I (RNA pol) (unstained spots), surrounded by POLGARF. (*E*) Coomassie blue stained SDS/PAGE with total cytoplasmic lysate and with nucleolus-enriched fractions of cells transfected with the indicated constructs. Protein bands corresponding to P32 and POLGARF were excised and subject to mass spectrometry.

of POLGARF-HiBiT, we detected a progressive increase in luciferase activity in HEK293T cell lysates. Interestingly, we also detected luciferase activity in the conditioned media (*SI Appendix*, Fig. S9). Extracellular HiBiT-containing protein was purified from the conditioned media with an engineered SNAP-LgBiT protein and SNAP magnetic beads. Upon fractionation by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS/PAGE), intracellular POLGARF-HiBiT migrates at 35 kDa as

expected (Fig. 5*A*) whereas the purified extracellular HiBiT fusion protein migrates at ∼17 kDa (Fig. 5*B*). LC-MS/MS analysis of the extracellular HiBiT fusion identified it as a heterogeneous population of C-terminal POLGARF fragments mainly produced by cleavages around positions 138 to 140 and 150 with a minor population of both longer and shorter proteoforms around the major cleavage positions (Fig. 5*C*). We called these fragments POLGARFin. Notably, POLGARFin fragments were not observed in cell lysates representing soluble cytosolic fractions, nor could we detect full-length POLGARF in the media (Fig. 5*A*). It seems unlikely that POLGARFin-HiBiT is released into the media from dead cells. The most probable explanation is that POLGARFin is secreted immediately after intracellular POLGARF cleavage. Alternatively, the full-length protein can be secreted and immediately cleaved outside of cells.

To investigate whether this secretion is regulated, we subjected POLGARF-HiBiT overexpressing cells to various treatments. It appeared that POLGARFin secretion is up-regulated by serum addition: When the media of transfected cells is supplemented with fresh media containing 10% fetal bovine serum (FBS), rapid accumulation of POLGARFin is observed within 30 min after stimulation with no further increase over time (Fig. 5*C*). In contrast, addition of serum-free media does not result in POLGARFin secretion. POLGARFin secretion in response to FBS-containing media does not depend on de novo protein synthesis as supplementation of cells with serum-rich media infused with cycloheximide does not prevent POLGARFin extracellular accumulation. Collectively, these data suggest that POLGARF is processed into actively secreted POLGARFin, which may be implicated in extracellular signaling events.

Discussion

The results presented above provide strong evidence for the unusual existence of alternative functional protein encoded within the *POLG* mRNA. This protein, termed POLGARF, is initiated by an unusually strong non-AUG codon due to its optimal nucleotide context (refs. 6 and 34, and this study), a 3′-terminal secondary structure that allows the initiating ribosome to stall (31), and the use of a CUG codon, the preferred non-AUG start codon (5, 6, 41, 43). When two different proteins are translated from a single mRNA, the efficiency of initiation at each start codon could set the ratio of product steady state levels, assuming each protein has similar stability. However, this ratio may vary during conditions in which initiation efficiency is altered. Eukaryotes have developed elaborate mechanisms for the recognition of the correct initiation codon, and the levels of certain initiation factors can regulate the fidelity of initiation, especially on suboptimal (near-cognate and AUG in poor context) start codons (2). While elevated levels of EIF1 can increase the stringency of start codon selection, elevated levels of EIF5 have the opposite effect. Here, we show that overexpression of either EIF1 or EIF5 had minimal effect on CUG-initiated POLGARF (Fig. 2*C*), indicating that this highly efficient CUG initiation is refractory to normal stringency controls. However, it should be noted that these findings are in contrast to the effect of eIF5 expression on GUG or CUG initiation from the *eIF4G2* context (6), suggesting that a 3′ stem loop (not present in *eIF4G2*) may be crucial for this effect.

Our analysis of the role of the conserved uORF in POLGARF initiation reveals that both leaky scanning and reinitiating ribosomes can equally start translation at the POLGARF CUG (*SI Appendix*, Fig. S4). However, since preventing translation of the uORF (therefore no reinitiation) results in predominantly CUG initiation then it follows that under normal conditions most of the ribosomes initiating the POLG ORF are reinitiating ribosomes. This raises the intriguing possibility that the ratio of POLG to POLGARF may be regulated by stress conditions.

Transposons are known to contribute to eukaryotic genome evolution. The activity of short interspersed nuclear element (SINE) transposons and their exonization made a particularly large impact on mammalian genomes (57). Here we propose that an event related to MIR transposition inflicted a birth of a dual-coding gene. How did POLGARF evolve, and why did it subsequently become fixed in placental mammals? All mammalian sequences, with the exception of platypus, share significant similarity with MIR sequences as determined with Dfam search

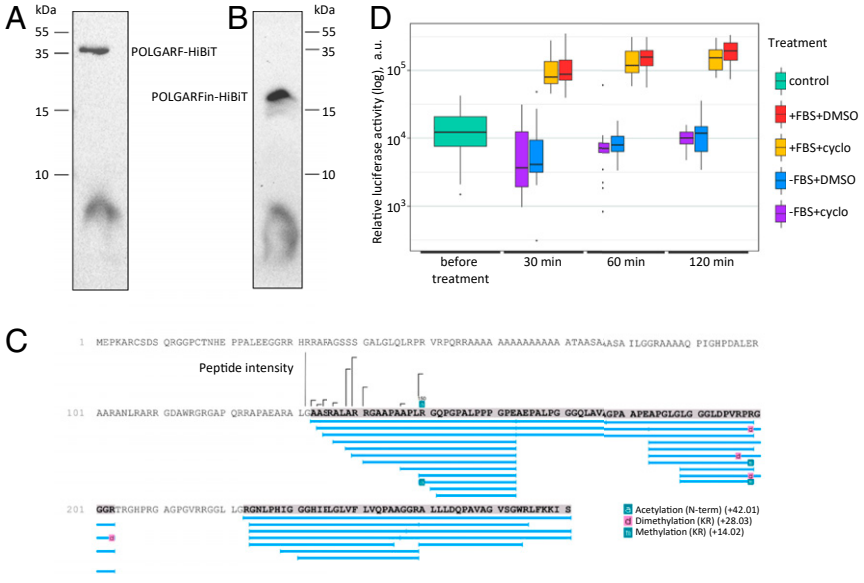

Fig. 5. POLGARF cleavage and secretion of its C-terminal fragment upon serum stimulation. (*A*) HiBiT blotting of cell lysates after expression of POLGARF-HiBiT. (*B*) HiBiT blotting of secreted HiBiT containing polypeptide purified from conditioned media from cells expressing POLGARF-HiBiT. (*C*) LC-MS sequencing of secreted POLGARFin-HiBiT. Peptide peak areas are shown for the sequences at the ragged N terminus of POLGARFin. Variable modifications include N-terminal acetylation and Lys and Arg (KR) methylation and dimethylation. (*D*) Measurement of HiBiT activity in the conditioned media from POLGARF-HiBiT transfected HEK293T cells. Twenty-four hours after transfection, the media was replaced with fresh DMEM +/− 10% FBS and +/− cycloheximide (cyclo); after indicated time points, the media aliquots were assayed with Nano-Glo HiBiT luciferase assay (n = 24).

(58). Therefore, it is likely that MIR insertion occurred after Monotremata diverged from the common ancestor of Marsupials and Placentals. We found that POLG sequences of many vertebrate species (including platypus) lack stop codons in one of the alternative frames in the first coding exon; thus, acquisition of an in-frame start codon could lead to the expression of the alternative frames. However, MIR did not contain a suitable start codon. Two subsequent mutations had to occur to enable POLGARF expression: a substitution of CCA with CTG and a single nucleotide deletion downstream that was necessary to place the CTG in-frame with the alternative ORF. Interestingly, these variants are common for both marsupials and placental mammals (shown in green in Fig. 3*B*), suggesting that proto-POLGARF existed in their common ancestor. However, marsupials have two variants (in red in the region shown in Fig. 3*B*) and several stop codons in the POLGARF frame further downstream, suggesting that POLGARF was subsequently lost in marsupials, while, in the common ancestor of placental mammals, it acquired a functional role and became fixed in subsequent lineages.

What may be the functional role of POLGARF? The polypeptide can be tentatively divided into four parts; notably, the 64-amino acid-long C terminus is the most conserved region, with 22 invariant amino acids (*SI Appendix*, Fig. S10). We failed to find any significant similarity between POLGARF and other known or predicted proteins or any similarity with known structural motifs. It seems likely that POLGARF is an intrinsically disordered protein (IDP) with a remarkably high isoelectric point (pI =12.05 for a human protein).

As a conserved IDP, POLGARF has a good potential to be an important regulatory protein. In cell signaling and regulation, IDPs emerged as parts of integrated circuits. Indeed, the capacity of IDPs to acquire numerous posttranslational modifications and change conformation in a context-dependent manner allows immense versatility of their interactomes (59, 60). The observed specific interaction of POLGARF with P32 as well as the modulatory effects of POLGARF on P32 localization and processing exemplify the importance of POLGARF for cell functioning (Fig. 4 and *SI Appendix*, Fig. S8). Along with P32, other putative interaction partners of POLGARF (TRIP13, CAMK2D, etc.) (*SI Appendix*, Table S1) can be considered as good candidates for mechanistic follow-up studies.

According to Riboseq analysis and reporter assays, the levels of POLG and POLGARF proteins should be comparable. However, according to proteomics data, in contrast to the moderately high levels of the housekeeping protein POLG, POLGARF's concentration is extremely low. Our discovery that POLGARFin can be secreted is likely the most probable explanation for this discrepancy. We propose that endogenous POLGARF may be almost completely processed and secreted outside of cells where it may participate in currently unknown cell-to-cell communication. We speculate that the levels of secreted POLGARFin reflect the capacity of the donor cell to have enough POLG to replicate mtDNA as both POLG and POLGARF are encoded in the same mRNA. In contrast, cells with decreased expression of *POLG* mRNA would also produce less POLGARFin.

Finally, these findings could have profound implications for the interpretation of POLG mutations. As a previously unknown dual coding gene, *POLG* could bear hidden mutations responsible for diseases of still unknown etiology. Among the known *POLG* mutations, some do not cause changes in the amino acid composition of POLG. Such synonymous single nucleotide variations (SNVs) are often considered harmless (except for their potential effects on splicing) (61). However, in dual coding regions (such as POLG/POLGARF described in this study), synonymous SNVs in one ORF are unlikely to be synonymous in the other ORF. This emphasizes the need to consider *POLGARF* variants in future studies.

Materials and Methods

Cell Culture. Here and elsewhere, all reagents were from Millipore-Sigma, unless stated otherwise. Human embryonic kidney HEK293T and Expi293F cells were from the American Type Culture Collection (Manassas, VA) and Thermo Fisher Scientific (Rockford, IL), respectively. HEK293T were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS, 2 mM L-glutamine, 100 U/mL penicillin/100 μg/mL streptomycin (complete DMEM) with or without 10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) (pH 7.2), in humidified atmosphere of 95% air and 5% CO_2 at 37 °C. Expi293F cells were maintained in Expi293 Expression medium (Thermo Fisher) at 8% CO_2 and 37 °C under continuous mixing at 160 RPM on an orbital shaker.

Plasmids and Constructs. To create plasmids for luciferase reporter assays, the POLG 5′ leader plus 33 nt of POLG CDS were amplified by PCR from HEK293T complementary DNA (cDNA) and cloned in dual luciferase vector p2-Luc (62) between HindIII and BamHI restriction sites. All other variants were generated by two-step PCR using appropriately designed mutagenic or nested deletion primers. For firefly luciferase reporters expressing varying N-terminal amino acids (Fig. 2 *B* and *C* and *SI Appendix*, Figs. S1 *C* and *D*, S2*B*, and S3*B*), HindIII and BamHI digested PCR amplicons were cloned into p2-LucSG—a modified variant of p2Luc encoding the porcine enterovirus StopGo sequence. Plasmids used to overexpress deregulated eIF1 (eIF1g*), and deregulated eIF5 (eIF5-AAA), have been described previously (43).

Plasmids for expression of POLGARF and P32 with various tags (FLAG and HiBiT at C terminus; GST, SNAP, and CLIP at N terminus) were created with modified pcDNA3.4 vector (Thermo Fisher), which contained a custom polylinker. SNAP and CLIP tags were originated from pSNAPf and pCLIPf vectors from New England Biolabs (NEB, Ipswich, MA).

To create the plasmid containing SNAP-LgBiT, the gene block containing the LgBiT sequence flanked by BamHI and NotI was ordered from IDT. The pcDNA3.4-SNAP-POLGARF plasmid was treated with BamHI and NotI, and the POLGARF coding fragment was exchanged with an LgBiT sequence to yield pcDNA3.4. SNAP-LgBiT. For more details about plasmid preparation, see *SI Appendix, Supplementary Methods*.

Cell Transfection. Four main transfection protocols were used (see *SI Appendix, Supplementary Methods* for details): one-day Lipofectamine 2000-based protocols for 1) suspended and 2) adherent HEK293T cells; 3) 40 to 48 h FuGENE protocol for adherent HEK293T cells; and 4) 48 h Expifectamine 293 protocol for expi293F cells.

Dual Luciferase Assay and Western Analysis. Firefly and Renilla luciferase activities were determined using the Dual Luciferase Stop & Glo System (Promega). Relative light units were measured on a Veritas Microplate Luminometer with two injectors (Turner Biosystems, Sunnyvale, CA). Transfected cells were lysed in 15 μL of 1× Passive Lysis Buffer (PLB; Promega), and light emission was measured following injection of 25 μL of either Renilla or Firefly luciferase substrate. Initiation efficiencies (% initiation) were determined by calculating relative luciferase activities (Firefly/Renilla) of test constructs and dividing by relative luciferase activities from replicate wells of control ATG constructs.

For Western blotting in Fig. 2, transfected cells were lysed in 100 μL of 1× PLB. Proteins were resolved by SDS/PAGE and transferred to Protran nitrocellulose membranes (GE Healthcare Life Sciences, Waukesha, WI), which were incubated at 4 °C overnight with primary antibodies. Anti-EIF1 was a kind gift from Ariel Stanhill (Technion-Israel Institute of Technology, Haifa, Israel). Immunoreactive bands were detected on membranes after incubation with appropriate fluorescently labeled secondary antibodies using a LI-COR Odyssey Infrared Imaging Scanner (LI-COR, Lincoln, NE).

For Western blotting analyses of POLGARF expression, transfected cells were washed with phosphate-buffered saline (PBS) and lysed for 20 min on ice with radioimmunoprecipitation assay (RIPA) buffer (Thermo Fisher), containing phosphatase and protease inhibitors; complete protease inhibitor mixture and PhosSTOP tablets were from Roche (Mannheim, Germany). After lysate clarification by centrifugation for 15 min at 14,000 × *g* and 4 °C, protein concentration was measured using BCA Protein Assay kit (Thermo Fisher) and equalized. Proteins were separated by 4 to 20% polyacrylamide gel electrophoresis using premade acrylamide gels and running buffers from GeneScript (Piscataway, NJ), transferred onto a 0.2-μm ImmobilonTM-P poly(vinylidene difluoride) (PVDF) membrane (Sigma) using the Hoefer TE 22 transfer system (Hoefer, Holliston, MA) and probed with antibodies

against POLGARF (1:1,000) and α-tubulin (1:5,000) in 5% fat-free milk in Tris buffered saline with Tween 20 (TBST) buffer (0.8% Tween 20). Immunoblots were analyzed using the Amersham ECL Prime Kit from GE Healthcare Life Sciences (Waukesha, WI) and the LAS-3000 Imager (Fujifilm). Quantitative image analysis was performed with the ImageJ program using α-tubulin signals for normalization.

GST and SNAP Pull Down Assays. For GST pull down assay, 2×10^8 expi293F cells transfected with pcDNA3.4-GST-POLGARF or with pcDNA3.4-GST were harvested by centrifugation and lysed in 2 mL of PLB. The lysates were diluted with PBS to 10 mL and incubated with 200 μL of GST-Sepharose (GE Healthcare) for 1 h on ice under agitation. GST resin was washed three times with 10 mL of PBS, and POLGARF bound proteins were eluted by incubation with 1 μL of Prescission Protease (GE Healthcare) in PBS at 4 °C overnight. The eluted proteins were resolved by SDS/PAGE and stained with Coomassie, and protein bands were excised and subjected to matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS/MS analysis.

For SNAP pull down assay, transfected cells were lysed in 200 μL of PLB supplemented with 1 mM DTT. The lysates were mixed with 30 μL of SNAP-magnetic beads (NEB) and incubated for 1 h at 24 °C on a thermomixer (900 rpm). After incubation, the beads were washed twice with 1 mL of PBS supplemented with 1 mM dithiothreitol (DTT), and bound proteins were eluted by boiling with SDS gel loading buffer. The samples were resolved on SDS/PAGE and immunoblotted with either anti-FLAG or with custom made anti-POLGARF antibodies.

Staining of Cells with Probes and Confocal Microscopy. Transfected cells were stained with SNAP-Cell TMR-Star or SNAP-Cell 647-SiR (NEB), both diluted 1:500 with complete DMEM for 30 min immediately prior to live cell imaging or immunostaining. Cell imaging was conducted on an Olympus FV1000 confocal laser scanning microscope with controlled CO_2, humidity, and temperature. Analysis was performed using FV1000 Viewer software (Olympus) and Microsoft Excel. See *SI Appendix, Supplementary Methods* for detailed protocols of live cell staining, immunostaining, and imaging.

Subcellular Fractionation. Nucleolar fraction was prepared from suspension expi293F cells according to the protocol described by Lamond and coauthors (63) with modifications. Cells (2×10^8) were harvested by centrifugation 48 h after transfection. Cell pellets were resuspended in 2 mL of hypotonic buffer A (10 mM Hepes [pH 7.9], 10 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, and Complete Protease Inhibitor tablet), incubated on ice for 15 min, homogenized 25 times in a Dounce tissue homogenizer (using a tight pestle "B"), and centrifuged at $1,000 \times g$ for 5 min. Pellets were resuspended in buffer A, incubated on ice for 10 min, and then homogenization followed by centrifugation was repeated in order to obtain a purer nuclear pellet. Then, the pellet was resuspended in 3 mL of S1 solution (0.25 M Sucrose, 10 mM $MgCl_2$, and Complete Protease Inhibitor tablet), layered over 3 mL of S2 solution (0.35 M Sucrose, 0.5 mM $MgCl_2$, and Complete Protease inhibitor tablet), and centrifuged at $1,430 \times g$ for 5 min. Purified nuclear pellets resuspended in 3 mL of S2 solution were sonicated 5× 10 s using SONICS Vibra cell (Sonics & Materials Inc., Newton, CT) (50% of max power). The sonicated sample was layered over 3 mL of S3 solution (0.88 M Sucrose, 0.5 mM $MgCl_2$, and Complete Protease Inhibitor tablet) and centrifuged at $3,000 \times g$ for 10 min. The resulting pellet contained the nucleolar enriched fraction.

MALDI-TOF Analysis. Protein spots were excised from the gel and digested with trypsin. Mass spectra were recorded on an UltrafleXtreme MALDI-TOF/TOF mass spectrometer (Bruker Daltonics) equipped with an Nd laser (354 nm). For the detailed protocol, see *SI Appendix, Supplementary Methods*.

Identification of POLGARF in Proteomics Data. Using the recently developed Post-Acquisition Targeted Searches technique (47), we submitted the eight longest predicted tryptic peptides from POLGARF for a targeted search in the preanalyzed HEK293 interactome dataset (48). The search yielded a total of 118 .raw file hits for 94 protein baits (*SI Appendix*, Table S1). For the baits with greater than two predicted peptide identifications, all .raw data were downloaded from the BioPlex2.0 website (http://bioplex.hms.harvard.edu) for in-house database search analysis.

POLGARF-predicted tryptic peptides were searched in PeptideAtlas (64). One of the peptides, AAAAQPIGHPDALER, turned out to be known and identified in several proteome datasets mostly connected with immune cells. In particular, this peptide was identified in a Jurkat dataset from ref. 51 and in immune cell analysis from ref. 52. For some other cell lines, including HEK293 cells, the dataset was taken from ref. 50. Corresponding datasets were downloaded from ProteomeXchange for in-house database search analysis. For the list of all publicly available datasets downloaded and reanalyzed against the POLGARF-containing protein database, see *SI Appendix*, Table S4.

To obtain reference peptide spectra, POLGARF was overexpressed in HEK293T and expi293F cells. The 293T cells were transfected in 12-well plates with 1 μg of pcDNA3.4 POLGARF or pcDNA3.4 Timer (as a negative control) with FuGENE transfection reagent according to the manufacturer's instructions. After 48 h, the cells were washed with PBS and lysed in a sonication bath in 150 μL of 1× Passive Lysis Buffer (Promega) supplemented with Protease Inhibitor mixture (Roche). Expi293F cells were transfected with pcDNA3.4 POLGARF or pcDNA3.4 (40 μg of DNA per 10^8 of cells) with Expifectamine 293 (Thermo Fisher). After 48 h, cells were washed with PBS and lysed in a sonication bath in 1× Passive Lysis Buffer (Promega) supplemented with Protease Inhibitor mixture (Roche). The lysates were heated for 10 min at 90 °C. Protein material was cleaned by precipitation, digested with trypsin, and analyzed by LC-MS. For the detailed protocol, see *SI Appendix, Supplementary Methods*.

LC-MS analysis was done on an Ultimate 3000 RSLCnano high-performance liquid chromatography (HPLC) system in a trap-elute configuration connected to a QExactive Plus mass spectrometer (Thermo Fisher). Peptides were separated by a 2-h gradient of acetonitrile in water with the addition of 0.1% formic acid (FA) in a home-packed 50 cm × 100 μm capillary column (65). MS data were collected in a Data-Dependent Acquisition (DDA) mode. Detailed parameters of the separation and detection are described in *SI Appendix, Supplementary Methods*.

Raw files from BioPlex, from ProteomeXchange, and the in-house generated LC-MS data were subjected to protein identification in Peaks Studio X (Bioinformatic Solution Inc., Waterloo, CA) against the UniProt *Homo sapiens* database containing both canonical and isoform proteoforms (version from 2019.08.26) with a manually attached POLGARF sequence. Search parameters included "trypsin with D|P" digestion with a maximum three miscleavages, precursor mass correction, 10 parts per million (ppm), and 0.05-Da error tolerance for precursor and fragment ions, respectively, oxidation (M) and deamidation (NQ) as variable modifications (maximum number of variable modification per peptide: five), and carbamidomethylation (C) as a fixed modification, Decoy–Fusion false discovery rate (FDR) estimation. Identification results were filtered by 0.1% peptide-spectrum match (PSM) FDR and one unique peptide per group, with the final protein FDR <1%.

Induction of POLGARFin Secretion. In initial experiments, HEK293T cells were transfected in three 4-well plates with POLGARF-HiBiT (FuGENE protocol). After every 24 h, a plate was taken out, medium and cells were harvested, and luminescence in cell lysates and in the media was measured using a Nano-Glo HiBiT luciferase assay. This experiment was repeated five times.

To investigate the conditions that mediate POLGARFin secretion, HEK293T cells were transfected with POLGARF-HiBiT in 48-well plate (FuGENE protocol). After 24 h posttransfection, medium was substituted for fresh DMEM with or without 10% FBS, with or without 100 μg/mL cycloheximide (or the same volume of dimethyl sulfoxide [DMSO] as control). Every 30 min after media exchange, aliquots of the media were collected, and HiBiT activity was analyzed with a Nano-Glo HiBiT luciferase assay (Promega). This experiment was repeated 21 times.

Purification of POLGARFin from Cultured Media and LC-MS Analysis. To prepare the bait protein SNAP-LgBiT containing lysate, pcDNA3.4 SNAP-LgBiT plasmid was transfected into expi293F cells (Expifectamine 293 protocol). Then, 48 h after transfection, 25×10^6 cells expressing SNAP-LgBiT were lysed in 5 mL of luciferase cell culture lysis (CCL) buffer (Promega). The lysate aliquots were stored at −80 °C.

To prepare the conditioned media for POLGARFin purification, expi293F cells were transfected with pcDNA3.4 POLGARF-HiBiT (and pcDNA3.4 for negative control, Expifectamine 293 protocol). Forty-eight hours after transfection, the media was substituted for a fresh one, and transfected cells were grown for an additional 48 h. Then, the cells were sedimented by centrifugation, and the conditioned media was used for POLGARFin purification.

First, SNAP-LgBiT protein was loaded onto SNAP magnetic beads. For that, 40 μL of prewashed SNAP magnetic beads were incubated with 500 μL of SNAP-LgBiT lysate for 1.5 h at room temperature. Then, the beads were washed in the CCL buffer supplemented with 0.5 M NaCl. Next, SNAP-LgBiT–modified magnetic beads were incubated with 10 mL of the conditioned medium for 30 min at room temperature with shaking. This procedure was repeated five times with fresh aliquots of the POLGARF-HiBiT–transfected cells medium. After the final incubation, the magnetic beads were washed in buffer A1000 (20 mM Tris·HCl, 10% glycerol, 1 mM DTT, ethylenediaminetetraacetic

acid [EDTA], 1 M KCl) and then washed two times in PBS (we have found that high salt wash does not disrupt association of HiBiT-containing proteins from SNAP-LgBiT). Bound proteins were eluted from the beads by incubation with 1x SDS-Loading dye for 5 min at 50 °C and resolved on Tris-Tricine SDS/PAGE. The gels were either stained with Coomassie (for mass spectrometry analysis) or transferred to a nitrocellulose membrane and then analyzed with HiBiT-blot. Briefly, the membrane was incubated in TBST buffer for 30 min, followed by sequential incubation with the buffer, LgBiT protein, and Furimazine from a Nano-Glo HiBiT luciferase assay (Promega) for 5 min followed by chemiluminescence detection on ChemiDoc XRS+ (Bio-Rad).

For LC-MS sequencing, the HiBiT-containing protein was purified from the conditioned media using the SNAP-LgBiT bait protein, resolved on a Tris-Tricine gel, and stained with Coomassie. The protein band corresponding to the mobility of the POLGARFin-HiBiT protein in the blot experiment of the same eluate was excised and analyzed by LC-MS/MS. In-gel protein digestion was done as in ref. 66 without protein reduction and Cys alkylation. The gel slice was divided into six pieces; three were digested with GluC, and the other three with GluC and trypsin. All samples were analyzed by LC-MS in the same way as for the full-cell proteome analysis described above. Peptides were separated on an Acclaim PepMap C18 2-μm column, 75 μm × 150 mm, using a 45-min gradient. Each sample was analyzed by a single LC-MS run.

The data were analyzed in PEAKS software against a Uniprot human database with a manually attached POLGARF sequence and the common contaminant database. The enzyme parameter was selected as "Specified by each sample" with semispecific digestion and chosen accordingly to the proteases used in the sample preparation. The results were filtered to PSM FDR 1%, which resulted in 2.1% peptide FDR and 0.2% protein group FDR.

Data Presentation. The number of biological replicates for each experiment is indicated in each figure legend. For all box plots, center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, and outliers are represented by dots. The alignment shown in Fig. 3*B* was manually refined.

Data Availability. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD016007 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD016007).

Note. While this manuscript was under revision, an independent study provided bioinformatics evidence for the existence of alternative reading frame translation in *POLG* (67).

ACKNOWLEDGMENTS. I.N.S., D.E.A., and M.S.M. acknowledge the Russian Science Foundation (RSF19-14-00152); G.L., P.B.F.O., A.M.M., and P.V.B. are funded by Science Foundation Ireland (SFI)-Health Research Board (HRB)-Wellcome Trust Biomedical Research Partnership (Investigator Award in Science 210692/Z/18/Z); S.J.K. is supported by the Irish Research Council. J.F.A. is funded by an Irish Research Council Advanced Laureate (IRCLA/2019/74); A.V.Z. and D.B.P. acknowledge SFI (SFI/12/RC/2276_P2). The MALDI-MS facility became available to us in the framework of the Moscow State University Development Program PNG 5.13. Funding for open access charge is from the Russian Science Foundation.

1. M. Kozak, Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. *Cell* **22**, 7–8 (1980).
2. A. G. Hinnebusch, The scanning mechanism of eukaryotic translation initiation. *Annu. Rev. Biochem.* **83**, 779–812 (2014).
3. A. M. Cigan, L. Feng, T. F. Donahue, tRNAi(met) functions in directing the scanning ribosome to the start site of translation. *Science* **242**, 93–97 (1988).
4. D. S. Peabody, Translation initiation at non-AUG triplets in mammalian cells. *J. Biol. Chem.* **264**, 5031–5035 (1989).
5. K. Asano, Why is start codon selection so precise in eukaryotes? *Translation (Austin)* **2**, e28387 (2014).
6. L. Tang *et al*., Competition between translation initiation factor eIF5 and its mimic protein 5MP determines non-AUG initiation rate genome-wide. *Nucleic Acids Res.* **45**, 11941–11953 (2017).
7. M. Kozak, Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**, 187–208 (1999).
8. W. L. Noderer *et al*., Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10**, 748 (2014).
9. N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, J. S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
10. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
11. S. Lee *et al*., Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2424–E2432 (2012).
12. C. Fritsch *et al*., Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22**, 2208–2218 (2012).
13. I. P. Ivanov, A. E. Firth, A. M. Michel, J. F. Atkins, P. V. Baranov, Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res.* **39**, 4220–4234 (2011).
14. P. Van Damme, D. Gawron, W. Van Criekinge, G. Menschaert, N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Mol. Cell. Proteomics* **13**, 1245–1261 (2014).
15. I. Tzani *et al*., Systematic analysis of the PTEN 5′ leader identifies a major AUU initiated proteoform. *Open Biol.* **6**, 150203 (2016).
16. A. G. Hinnebusch, I. P. Ivanov, N. Sonenberg, Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
17. T. G. Johnstone, A. A. Bazzini, A. J. Giraldez, Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35**, 706–723 (2016).
18. S. E. Calvo, D. J. Pagliarini, V. K. Mootha, Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 7507–7512 (2009).
19. T. E. Dever *et al*., Phosphorylation of initiation factor 2 alpha by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell* **68**, 585–596 (1992).
20. D. E. Andreev *et al*., Translation of 5′ leaders is pervasive in genes resistant to eIF2 repression. *eLife* **4**, e03971 (2015).
21. K. M. Vattem, R. C. Wek, Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11269–11274 (2004).
22. C. Sidrauski, A. M. McGeachy, N. T. Ingolia, P. Walter, The small molecule ISRIB reverses the effects of eIF2α phosphorylation on translation and stress granule assembly. *eLife* **4**, e05033 (2015).
23. A. Wiese, N. Elzinga, B. Wobbes, S. Smeekens, A conserved upstream open reading frame mediates sucrose-induced repression of translation. *Plant Cell* **16**, 1717–1729 (2004).
24. J. R. Hill, D. R. Morris, Cell-specific translational regulation of S-adenosylmethionine decarboxylase mRNA. Dependence on translation and coding capacity of the cis-acting upstream open reading frame. *J. Biol. Chem.* **268**, 726–731 (1993).
25. I. P. Ivanov *et al*., Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. *Mol. Cell* **70**, 254–264.e6 (2018).
26. I. P. Ivanov, G. Loughran, J. F. Atkins, uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10079–10084 (2008).
27. A. Trifunovic *et al*., Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* **429**, 417–423 (2004).
28. A. M. Michel *et al*., GWIPS-viz: Development of a ribo-seq genome browser. *Nucleic Acids Res.* **42**, D859–D864 (2014).
29. S. J. Kiniry, P. B. F. O'Connor, A. M. Michel, P. V. Baranov, Trips-Viz: A transcriptome browser for exploring ribo-seq data. *Nucleic Acids Res.* **47**, D847–D852 (2019).
30. S. E. Kolitz, J. E. Takacs, J. R. Lorsch, Kinetic and thermodynamic analysis of the role of start codon/anticodon base pairing during eukaryotic translation initiation. *RNA* **15**, 138–152 (2009).
31. M. Kozak, Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8301–8305 (1990).
32. M. Kozak, Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* **266**, 19867–19870 (1991).
33. H. M. Martinez, Detecting pseudoknots and other local base-pairing structures in RNA sequences. *Methods Enzymol.* **183**, 306–317 (1990).
34. A. J. Diaz de Arce, W. L. Noderer, C. L. Wang, Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res.* **46**, 985–994 (2018).
35. R. Boeck, D. Kolakofsky, Positions +5 and +6 can be major determinants of the efficiency of non-AUG initiation codons for protein synthesis. *EMBO J.* **13**, 3608–3617 (1994).
36. S. Grünert, R. J. Jackson, The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons. *EMBO J.* **13**, 3618–3630 (1994).
37. H. Imataka, H. S. Olsen, N. Sonenberg, A new translational regulator with homology to eukaryotic translation initiation factor 4G. *EMBO J.* **16**, 817–825 (1997).
38. J. H. Xiao, I. Davidson, H. Matthes, J. M. Garnier, P. Chambon, Cloning, expression, and transcriptional properties of the human enhancer factor TEF-1. *Cell* **65**, 551–568 (1991).
39. G. Loughran, A. E. Firth, J. F. Atkins, I. P. Ivanov, Translational autoregulation of BZW1 and BZW2 expression by modulating the stringency of start codon selection. *PLoS One* **13**, e0192648 (2018).
40. A. G. Hinnebusch, Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol. Mol. Biol. Rev.* **75**, 434–467 (2011).
41. G. Loughran, M. S. Sachs, J. F. Atkins, I. P. Ivanov, Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. *Nucleic Acids Res.* **40**, 2898–2906 (2012).
42. C. Kozel *et al*., Overexpression of eIF5 or its protein mimic 5MP perturbs eIF2 function and induces ATF4 translation through delayed re-initiation. *Nucleic Acids Res.* **44**, 8704–8713 (2016).

43. I. P. Ivanov, G. Loughran, M. S. Sachs, J. F. Atkins, Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18056–18060 (2010).
44. M. Kozak, Constraints on reinitiation of translation in mammals. *Nucleic Acids Res.* **29**, 5226–5232 (2001).
45. K. R. Rosenbloom *et al*., The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
46. M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
47. Y. Gao, J. Ma, A. Saghatelian, J. R. I. Yates, Targeted searches for novel peptides in big mass spectrometry data sets. https://doi.org/10.1101/239863 (25 December 2017).
48. E. L. Huttlin *et al*., Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).
49. M. Johnson *et al*., NCBI BLAST: A better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
50. D. B. Bekker-Jensen *et al*., An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst.* **4**, 587–599.e4 (2017).
51. N. A. Kulak, P. E. Geyer, M. Mann, Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteomics* **16**, 694–705 (2017).
52. J. C. Rieckmann *et al*., Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat. Immunol.* **18**, 583–593 (2017).
53. J. Jiang, Y. Zhang, A. R. Krainer, R. M. Xu, Crystal structure of human p32, a doughnut-shaped acidic mitochondrial matrix protein. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3572–3577 (1999).
54. V. Fogal *et al*., Mitochondrial p32 protein is a critical regulator of tumor metabolism via maintenance of oxidative phosphorylation. *Mol. Cell. Biol.* **30**, 1303–1318 (2010).
55. M. Yagi *et al*., p32/gC1qR is indispensable for fetal development and mitochondrial translation: Importance of its RNA-binding ability. *Nucleic Acids Res.* **40**, 9717–9737 (2012).
56. M. K. Schwinn *et al*., CRISPR-mediated tagging of endogenous proteins with a luminescent peptide. *ACS Chem. Biol.* **13**, 467–474 (2018).
57. J. Schmitz, SINEs as driving forces in genome evolution. *Genome Dyn.* **7**, 92–107 (2012).
58. R. Hubley *et al*., The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
59. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
60. R. B. Berlow, H. J. Dyson, P. E. Wright, Expanding the paradigm: Intrinsically disordered proteins and allosteric regulation. *J. Mol. Biol.* **430**, 2309–2320 (2018).
61. S. Richards *et al*., Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
62. G. Grentzmann, J. A. Ingram, P. J. Kelly, R. F. Gesteland, J. F. Atkins, A dual-luciferase reporter system for studying recoding signals. *RNA* **4**, 479–486 (1998).
63. D. Bensaddek, A. Nicolas, A. I. Lamond, Quantitative proteomic analysis of the human nucleolus. *Methods Mol. Biol.* **1455**, 249–262 (2016).
64. F. Desiere *et al*., The PeptideAtlas project. *Nucleic Acids Res.* **34**, D655–D658 (2006).
65. S. I. Kovalchuk, O. N. Jensen, A. Rogowska-Wrzesinska, FlashPack: Fast and simple preparation of ultrahigh-performance capillary columns for LC-MS. *Mol. Cell. Proteomics* **18**, 383–390 (2019).
66. A. Shevchenko, H. Tomas, J. Havlis, J. V. Olsen, M. Mann, In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2006).
67. Y. A. Khan *et al*., Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. *BMC Genet.* **21**, 25 (2020).
68. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).