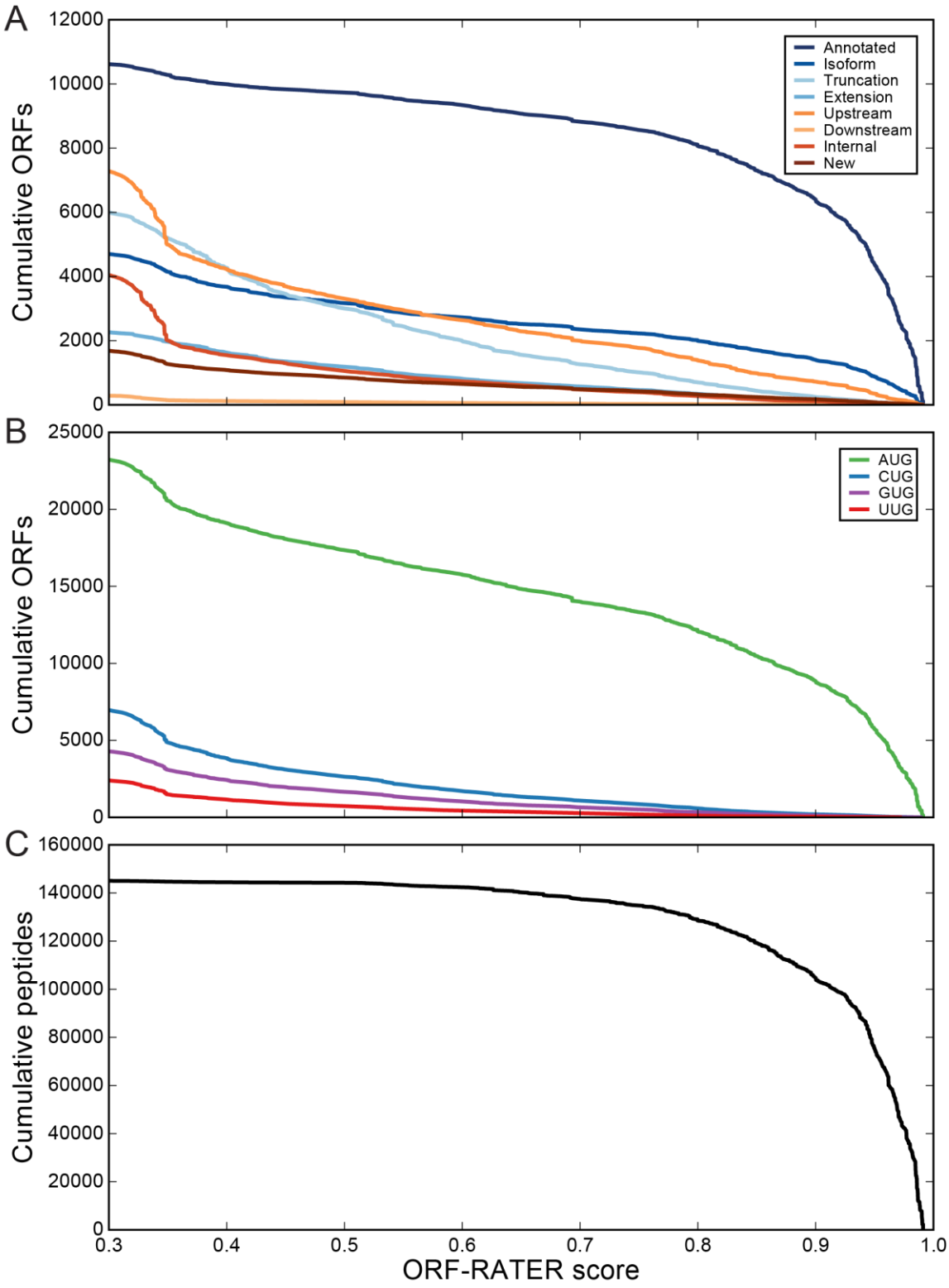

A
Cumulative ORFs
12000
10000
8000
6000
4000
2000
0
Annotated
Isoform
Truncation
Extension
Upstream
Downstream
Internal
New
B
Cumulative ORFs
25000
20000
15000
10000
5000
0
AUG
CUG
GUG
UUG
C
Cumulative peptides
160000
140000
120000
100000
80000
60000
40000
20000
0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ORF-RATER score

Figure S1. Full distributions of ORF-RATER scores, related to Figures 1 and 2 and Table 1.

(**A**) For each type of ORF, the cumulative distribution of the ORF-RATER score (i.e., the number of ORFs receiving at least that score) is plotted. (**B**) The cumulative distribution of the ORF-RATER score is plotted for each of the four codon species considered as potential translation initiation sites. (**C**) The cumulative number of peptides captured by ORF-RATER is plotted as a function of score threshold. If peptides could match multiple CDSs, the one with the highest ORF-RATER score was selected. In addition to the 145,033 peptides documented in the plot, an additional 4,074 (for 149,107 total) were identified that matched only CDSs not considered by ORF-RATER (e.g. due to low sequencing coverage) but present in the UniProt mouse database. High-confidence translated ORFs were defined as those receiving scores in the range 0.8–1, a conservative threshold intended to limit the number of false positive identifications. Some bona fide translated CDSs are not detected at this threshold, due to a low translation rate or ambiguous ORF structure on their transcript(s) (e.g. multiple neighboring in-frame AUGs). Additionally, some MS-detectable proteins may be missed by ribosome profiling if they are no longer being translated.

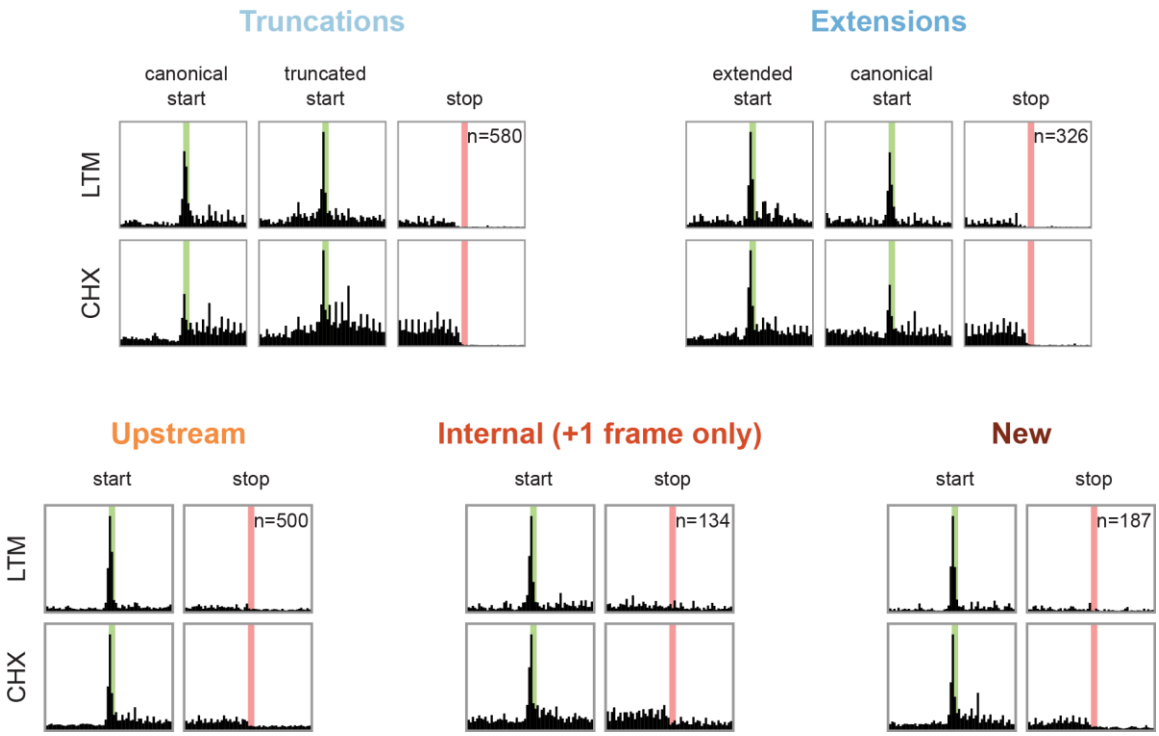


Figure S2. Metagene profiles for novel CDSs following LTM or CHX treatment, related to Figure 2.

Metagene profiles of each class of new CDS following treatment with LTM or CHX display hallmarks of translation, such as peaks of density at translation initiation sites in both datasets, and elevated average density in the body of CDSs following CHX treatment.

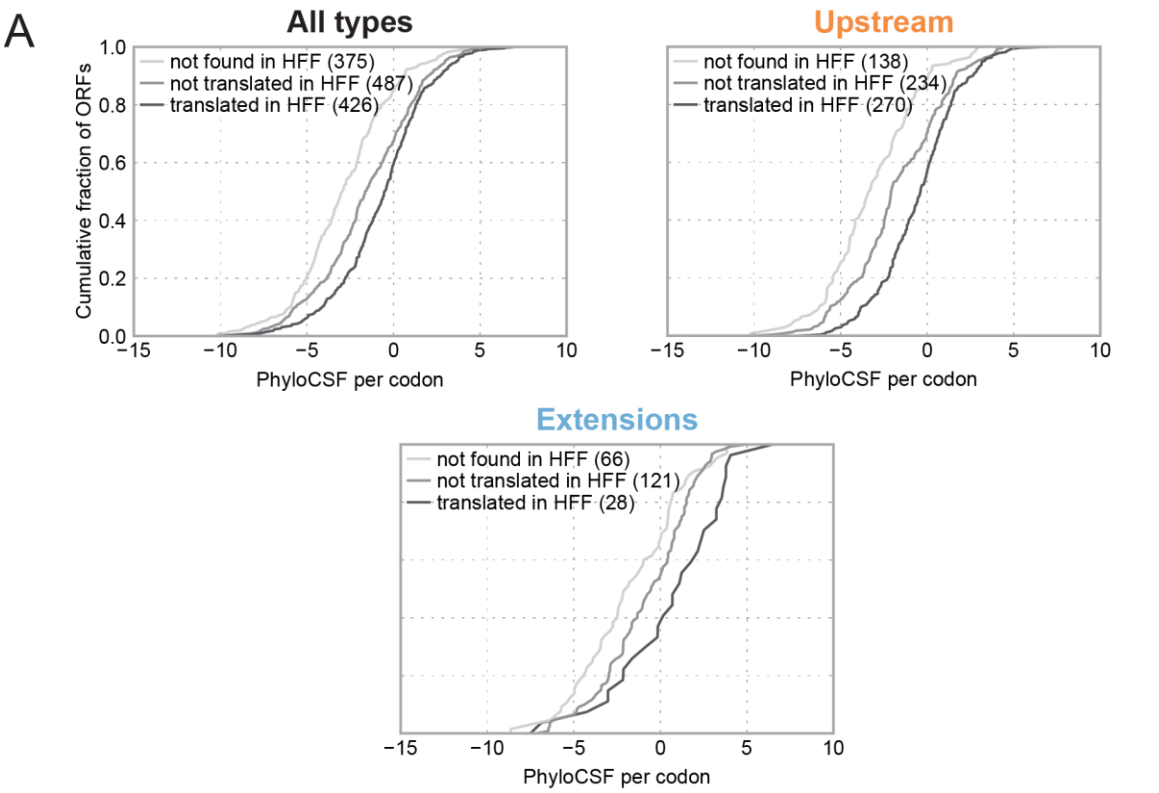
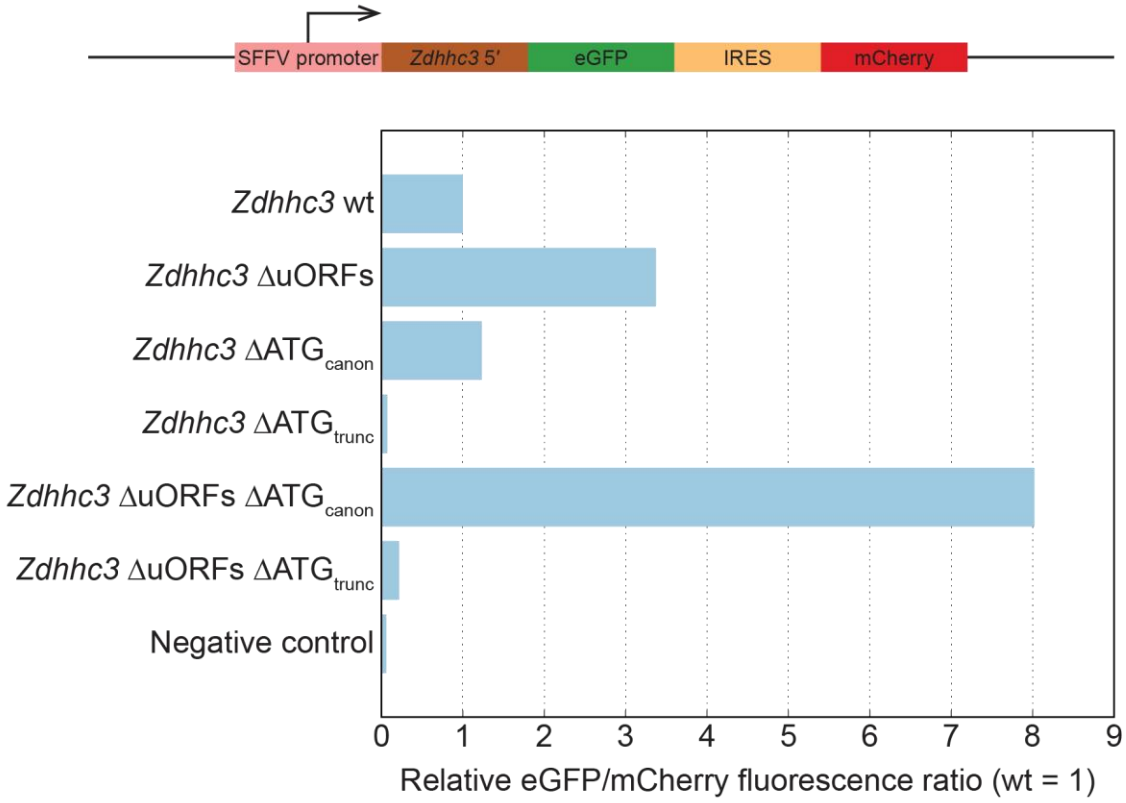


Figure S3. A significant subset of CDSs expressed in both mouse BMDCs and HFFs do not appear to be conserved for protein sequence, related to Figures 5 and 6.
(**A**) Cumulative distributions of per-codon PhyloCSF scores for the non-annotated portions of all novel CDSs (left), uORFs (center), and extensions (right) identified as translated in mouse BMDCs. CDSs for which no homologous translation initiation site was identified in the HFF transcriptome received lower PhyloCSF scores; those for which a homologous site was found in HFFs but not identified as translated received intermediate PhyloCSF scores; and those for which the homologous site was identified as translated in HFFs received greater PhyloCSF scores. Nonetheless, a significant fraction of the CDSs translated in both mouse BMDCs and HFFs received negative PhyloCSF scores, suggesting that the translation of those CDSs is conserved independent of the encoded polypeptide sequence. (**B**) A uORF of *Zdhhc3* encodes a peptide whose sequence does not appear to be strongly conserved among Euarchontoglires, despite being translated in both mouse BMDCs and HFFs. This uORF received a PhyloCSF score of −101 decibans (−1.86 decibans/codon).

A


SFFV promoter
Zdhhc3 5'
eGFP
IRES
mCherry
Zdhhc3 wt
Zdhhc3 ΔuORFs
Zdhhc3 ΔATG canon
Zdhhc3 ΔATG trunc
Zdhhc3 ΔuORFs ΔATG canon
Zdhhc3 ΔuORFs ΔATG trunc
Negative control
0
1
2
3
4
5
6
7
8
9
Relative eGFP/mCherry fluorescence ratio (wt = 1)

B

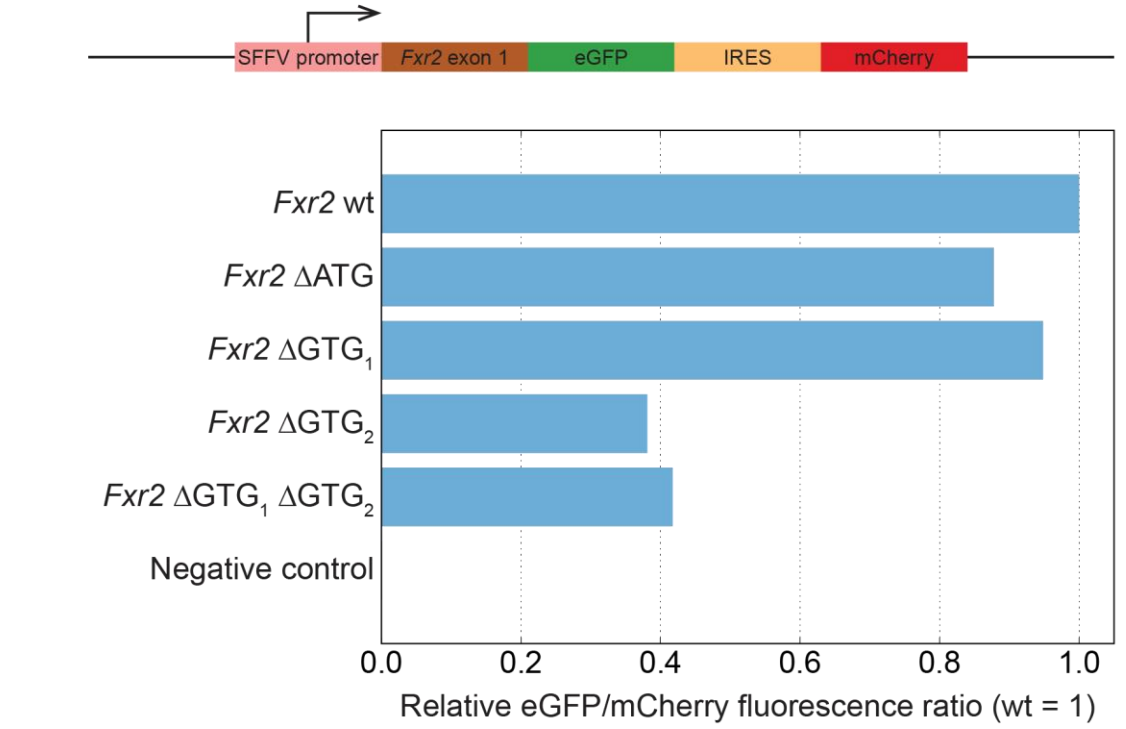

SFFV promoter
Fxr2 exon 1
eGFP
IRES
mCherry
Fxr2 wt
Fxr2 ΔATG
Fxr2 ΔGTG 1
Fxr2 ΔGTG 2
Fxr2 ΔGTG 1 ΔGTG 2
Negative control
0.0
0.2
0.4
0.6
0.8
1.0
Relative eGFP/mCherry fluorescence ratio (wt = 1)

Figure S4. Fluorescent reporter constructs confirm translated CDSs of *Zdhhc3* and *Fxr2*, related to Figures 5 and 6.

(**A**) Top, *Zdhhc3* reporter construct design. The 5′ portion of the *Zdhhc3* mRNA (ending immediately prior to the first annotated transmembrane domain) was cloned from human cDNA and fused to an in-frame eGFP sequence lacking its initial AUG. To enable normalization for transfection and transcription efficiency, an mCherry sequence whose translation was driven by an internal ribosome entry site (IRES) was included downstream of the fused eGFP CDS. Point mutants of the *Zdhhc3* gene sequence corresponding to removal of all three of the uORFs, removal of the canonical translation start site, and/or removal of the truncated translation start site were also generated (see Methods for detail). Bottom, fluorescence measurements following transient transfection of *Zdhhc3* reporter constructs into HEK 293T cells. Removal of the uORFs increased eGFP fluorescence, suggesting that translation of the uORFs inhibits downstream translation. Removal of the canonical ATG increased eGFP fluorescence marginally, whereas removal of the truncated translation start site abolished eGFP fluorescence, suggesting that essentially all translation of the *Zdhhc3* protein initiates at the downstream ATG. Results for constructs lacking the uORF translation start sites in addition to either the canonical or truncated ATGs suggest that even in the absence of uORFs, the truncated ATG remains the preferred translation initiation site. (**B**) Top, *Fxr2* reporter construct design. The reporters were identical to those constructed for *Zdhhc3* (above) except that the sequence fused to eGFP was cloned from the human genomic sequence for the first exon of *Fxr2*. The eGFP fluorescence level is most significantly reduced following removal of the second of two upstream in-frame GTG codons identified as a translation start site in mouse or human cells.

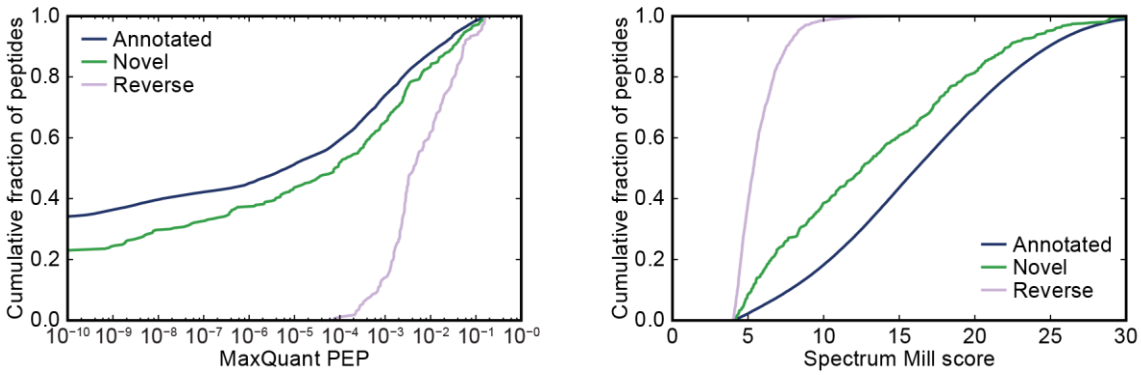


Figure S5. Peptides assigned to novel CDSs receive reasonable identification scores, related to Figure 6.

Tryptic peptides that match only newly identified CDSs score slightly more poorly than peptides matching annotated proteins, but are much more reliably identified than those matching reversed (decoy) peptides, both for MaxQuant (left) and Spectrum Mill (right). Note that lower MaxQuant PEP scores (left) indicate a more-reliable identification, whereas lower Spectrum Mill scores (right) indicate a less-reliable identification.

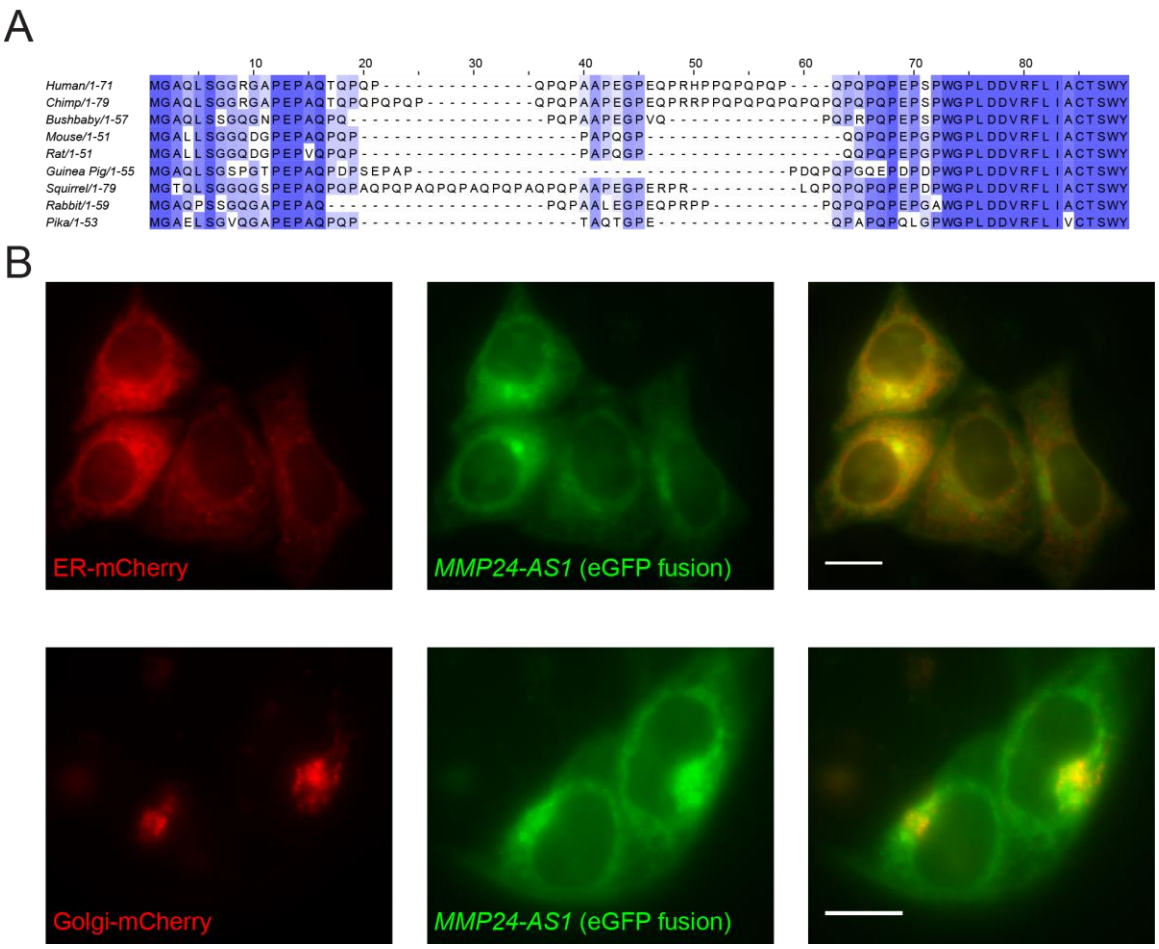


Figure S6. *BC029722* encodes a conserved protein that localizes to the ER and Golgi apparatus, related to Figure 6.

(**A**) A multiple sequence alignment of the protein encoded by the *BC029722* gene indicates that although the length of the protein varies, its N- and C-terminal regions are well conserved among the Euarchontoglires. A homologous protein was not identified in Rhesus macaque. (**B**) The human *BC029722* homologue *MMP24-AS1* fused to eGFP at its C-terminus colocalizes with ER (top) and Golgi (bottom) markers when coexpressed in HeLa cells. Scale bars, 10 μm.

Table S1. High-confidence translated ORFs identified by ORF-RATER in mouse BMDCs, related to Figures 2 and 4.

The 13,075 high-confidence translated CDSs ORF-RATER identifies in mouse BMDCs are listed in this table. Genomic coordinates refer to the mm9 build of the mouse genome. The "Homologous codon in HFF?" column indicates whether a corresponding NUG codon could be identified in the HFF transcriptome; the "Translated codon in HFF?" column indicates whether a high-confidence translation event was identified in HFFs that initiated at that codon. "Novel peptides" indicates the number of peptides that could be assigned to that CDS after excluding those that could have arisen from annotated proteins. "PhyloCSF" is the direct output from the PhyloCSF algorithm (measured in decibans) for only the portion(s) of the CDS non-overlapping with annotated CDSs, as detailed in Methods. "PhyloCSF number of codons" indicates the size of the region to which the PhyloCSF algorithm was applied. The final 26 columns are RPKM values for each ORF at each time point before or during LPS stimulation. For these columns, "CHX1" and "CHX2" refer to the two CHX-treated replicates, whereas "ND" was collected in the absence of added drugs. The second part of the column name indicates the number of hours of LPS treatment (e.g. 0h means untreated, 0.5h means half an hour of treatment). Some ORFs could not be quantified due to their being too short. RPKM values for very short or highly overlapping ORFs may be unreliable.

Table S2. Shotgun proteomics identifies peptides corresponding to some novel translated CDSs, related to Figure 6

	High-confidence set[a]		Extended set[b]	
	Proteins	Peptides	Proteins	Peptides
Extensions	40	95	36	41
Isoforms	74	159	39	42
New	7	29	6	19
Truncations	38	46	76	87
Upstream	6	21	2	2
Total	165	350	159	191

a The high-confidence set includes ORFs assigned scores of 0.8–1 by ORF-RATER
b The extended set includes ORFs assigned scores of 0.5–0.8 by ORF-RATER

Table S3. Peptides identified from mouse DCs with their most parsimonious protein source, related to Figure 6.

The 149,107 tryptic peptides detected from mouse DCs are listed in this table. The first column lists each peptide's amino acid sequence. "MaxQuant score" and "MaxQuant PEP" are the "Score" and "PEP" values reported by MaxQuant for those peptides it identified. "Spectrum mill score" is the "score" reported by Spectrum Mill for those peptides it identified. Not all peptides were identified by both programs. The following six columns ("Gene", "Chromosome", "Start codon coordinate", "Stop codon coordinate", "Strand", "Protein length (AA)", and "ORF type") are as in Table S1. "Only novel matches?" indicates whether the peptide could only have arisen from proteins newly identified by ORF-RATER; a value of "FALSE" indicates that the peptide could match an entry in UniProt, Ensembl, and/or UCSC KnownGene, even if this is not the most parsimonious matching entry.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

BMDCs growth conditions

All animal protocols were reviewed and approved by the MIT/Whitehead Institute/Broad Institute Committee on Animal Care (CAC protocol 0609-058-12). To obtain sufficient number of cells, we implemented a version of the BMDC isolation protocol as described previously (Jovanovic et al., 2015). Briefly, 6–8 week old female C57BL/6J mice were obtained from the Jackson Laboratories. RPMI medium (Invitrogen) supplemented with 10% heat inactivated FBS (Invitrogen), β-mercaptoethanol (50 μM, Invitrogen), L-glutamine (2 mM, VWR), penicillin/streptomycin (100 U/mL, VWR), MEM non-essential amino acids (1X, VWR), HEPES (10 mM, VWR), sodium pyruvate (1 mM, VWR), and GM-CSF (20 ng/mL; Peprotech) was used throughout the study. At day 0, BMDCs were collected from femora and tibiae and plated on twenty (per mouse), 100 mm non-tissue culture-treated plastic dishes using 10 mL medium per plate. At day 2, cells were fed with another 10 mL medium per dish. At day 5, cells were harvested from 15 mL of the supernatant by spinning at 1,400 rpm for 5 minutes; pellets were resuspended with 5 mL medium and added back to the original dish. Cells were fed with another 5 mL medium at day 7. At day 8, all non-adherent and loosely bound cells were collected and harvested by centrifugation. Cells were then resuspended with medium, plated at a concentration of 10×10^6 cells in 10 mL medium per 100 mm dish. At day 9, cells were stimulated for various time points with LPS (100 ng/mL, rough, ultrapure *E. coli* K12 strain, Invitrogen).

BMDC treatment for ribosome profiling

We generated two independent LPS time courses (9 time points: 0, 0.5, 1, 2, 4, 6, 8, 9, and 12 hours) and one Mock treated (LPS-free media) time course (same time points) for BMDCs to be treated with CHX. We generated one LPS time course (same time points) for ND treatment. The LTM-treated sample was an equal mix of BMDCs stimulated with LPS at 8 of the 9 time points (all except 0.5-hour). Harr-treated samples were unstimulated BMDCs (0-hour) and an equal mix of BMDCs stimulated with LPS for 2, 4, 6, and 9 hours. We used 20 million cells for each single time point sample or pooled sample.

For Harr inhibition of initiation, BMDCs were treated with 1 μg/mL Harr (LKT Laboratories) at 37°C for 5 min. For LTM inhibition of initiation, BMDCs were treated with 50 μg/mL LTM (Millipore) at 37°C for 30 min. For CHX inhibition of elongation, BMDCs were treated with 100 μg/mL CHX (Sigma) at 37°C for 1 min. For all three inhibitor treatments, cells were then treated with 100 μg/ml CHX at 37°C for 1 min, and then washed twice with cold PBS containing 100 μg/mL CHX. For ND treatment, media was removed from BMDCs, followed by flash freezing of the cells in liquid nitrogen. For all of the treatments, cells were next lysed by triturating 10 times with 400 μL lysis buffer (20 mM Tris pH 7.56, 150 mM NaCl, 5 mM $MgCl_2$, 1% Triton X-100 [Sigma], 1 mM DTT [Sigma], 8% glycerol, 100 μg/mL CHX, 12 units/mL Turbo DNase [Life Technologies]). Lysates were clarified by centrifuging at 20,000×g for 10 min at 4°C. Clarified lysate was flash frozen in liquid nitrogen.

BMDC protein isolation and processing for subsequent mass spectrometry

In order to capture proteins expressed at different times before or during LPS stimulation, we collected 4 samples over the time course (0, 2, 6, and 12 hours). After stimulation for the appropriate time, cells were washed twice with PBS and lysed for 30 min in ice-cold lysis urea buffer (8 M urea; 75 mM NaCl, 50 mM Tris HCl pH 8.0, 1 mM EDTA, 2 μg/mL aprotinin [Sigma, A6103], 10 μg/mL leupeptin [Roche, #11017101001], 1 mM PMSF [Sigma, 78830]). Lysates were centrifuged at 20,000×g for 10 min, and protein concentrations of the clarified lysates were measured via BCA assay (Pierce). 100 μg of total protein per time point were treated for 45 min with 5 mM dithiothreitol (Thermo Scientific, 20291) to reduce protein disulfide bonds and alkylated for 45 min with 10 mM iodoacetamide. Samples were then diluted 1:4 with 50 mM Tris HCl, pH 8.0, to reduce the urea concentration to < 2 M. Lysates were digested overnight at room temperature with trypsin in a 1:50 enzyme-to-substrate ratio (Promega, V511X) on a shaker. Tryptic peptides were desalted on C18 StageTips (Rappsilber et al., 2007) and evaporated to dryness in a vacuum concentrator.

Desalted peptides were labeled with the iTRAQ reagent according to the manufacturer's instructions (AB Sciex) and as previously described (Mertins et al., 2012). Briefly, 1 unit of iTRAQ reagent was used per sample (time point). 100 μg of peptides were dissolved in 30 μL of 0.5 M TEAB pH 8.5 solution and the iTRAQ reagent was added in 70 μL of ethanol. After 1 hour incubation the reaction was stopped with 50 mM Tris/HCl (pH 8.0). Differentially labeled peptides were mixed and subsequently desalted on C18 StageTips (Rappsilber et al., 2007) and evaporated to dryness in a vacuum concentrator.

To reduce peptide complexity and achieve deeper proteome coverage, samples were separated by basic reversed-phase chromatography as previously described (Mertins et al., 2013). Briefly, desalted peptides were reconstituted in 900 μL 20 mM ammonium formate, pH 10, and centrifuged at 10,000×g to clarify the mixture before it was transferred into autosampler tubes. Basic reversed-phase chromatography was conducted on a Zorbax 300 Å Extend-C18 column, using an Agilent 1100 Series HPLC instrument. The separations were performed on a 2.1 mm × 150 mm column (Agilent, 3.5 μm bead size). Prior to each separation, columns were monitored for efficient separation with standard mixtures containing 6 peptides. Solvent A (2% acetonitrile, 5 mM ammonium formate, pH 10), and a nonlinear increasing concentration of solvent B (90% acetonitrile, 5 mM ammonium formate, pH 10) were used to separate peptides by their hydrophobicity at a high pH. We used a flow rate of 0.2 mL/min and increased the percentage of solvent B in a nonlinear gradient with 4 different slopes (0% for 1 min; 0% to 9% in 6 min; 9% to 13% in 8 min; 13% to 28.5% in 46.5 min; 28.5% to 34% in 5.5 min; 34% to 60% in 23 min; 60% for 26 min). Eluted peptides were collected in 96-well plates with 1 min (= 0.2 mL) fractions. Early eluting peptides were collected in fraction "A", which is a combined sample of all fractions collected before any major UV-214 signals were detected. The peptide samples were combined into 24 to be used for proteome analysis. Subfractions were achieved in a serpentine, concatenated pattern, combining eluted fractions from the beginning, middle, and end of the run to generate subfractions of similar complexities that contain hydrophilic as well as hydrophobic peptides. For high-scale proteome analysis every 24th fraction was combined (1+25+49; 2+26+50; etc.). Subfractions were acidified to a final concentration of 1% formic acid and desalted on

C18 StageTips (Rappsilber et al., 2007). LC-MS/MS analysis was performed as previously described (Mertins et al., 2013). Each of the 24 subfractions was analyzed twice on the LC-MS/MS in order to increase proteome coverage.

Ribosome profiling

Ribosome-protected footprints were prepared for sequencing as previously described (Stern-Ginossar et al., 2012). Briefly, clarified cell lysates were treated with RNase I (Life Technologies) to digest RNA not protected by ribosomes. High molecular weight species were isolated by centrifuging lysates through a 34% sucrose cushion at 200,000×g for 4 hours at 4°C. Pellets were resuspended in Trizol (Life Technologies) and the RNA fraction isolated per the manufacturer's instructions. LTM and ND samples were treated with RiboZero (Epicentre Biotechnologies) after Trizol extraction, whereas Harr and CHX samples were depleted of rRNA by subtractive hybridization. Total RNase I-treated RNA was size-separated by gel electrophoresis, and RNAs of size 28–34 nucleotides were purified. These ribosome protected fragments were cloned and sequenced via a single-end run on an Illumina HiSeq 2500 sequencer.

Transcriptome mapping of reads

Sequenced reads were first stripped of linker sequences used in the cloning procedure. The resultant sequences were then filtered with Bowtie2 in --local mode for a collection of rRNA, snoRNA, snRNA, miRNA, tRNA, mitochondrial rRNA, and mitochondrial tRNA sequences compiled from UCSC and Ensembl annotations. The remaining reads were then aligned with Tophat to a BMDC-specific transcriptome, which was assembled from previously acquired RNA sequencing data (Shalek et al., 2013) using the PASA platform (Haas et al., 2003). Reads not mapping to the BMDC-specific transcriptome were then mapped with Tophat to the set of transcripts in UCSC Known Gene (version of March 6, 2014) and Ensembl (version of March 11, 2014). The following settings were used for Tophat alignments: --b2-very-sensitive --transcriptome-only --no-novel-juncs --max-multihits=64.

P-site assignment

Reads whose mapped positions overlapped annotated start codons were grouped by drug treatment, read length, and presence of a mismatch at their 5′-most nucleotide (reads with more than one such mismatch were excluded). For each of these groups, the most frequent relative position of the adenine of the annotated start codon was assigned as the P-site offset. The relevant P-site offset was then applied to all transcriptome-mapped reads so that each read is mapped to a single transcript position.

Transcript selection

Transcripts were removed from the collection if they included no mapped reads or if the positions to which reads were assigned were entirely contained within a shorter transcript. For each transcript, "multimapping" 29-nt subsequences identical to subsequences on other transcripts at other genomic positions were identified, as well as the number of 29- or 30-nt reads (in the ND dataset for the BMDCs, the CHX dataset for the HFFs) aligning to each position. Transcripts were required to have at least 64 reads to be kept, and transcripts for which 1/5 or more of the reads aligned to a single position

were also excluded. Transcripts overlapping annotated pseudogenes from Ensembl/GENCODE (Pei et al., 2012) were excluded if more than 1/3 of their aligned reads were in multimapping positions. Finally, the number of reads multimapping among the remaining transcripts was calculated, and transcripts were removed if their fraction of multimapped reads exceeded their fraction of multimapped positions by a value greater than 1/3 (e.g., a transcript with 50% multimapping positions and 50% multimapping reads would be kept, but one with 15% multimapping positions and 50% multimapping reads would be removed).

ORF-RATER

Sets of transcripts sharing genomic positions on the same strand (i.e., RNA isoforms) were grouped into "transcript families". Reads aligning to the transcripts within these families were assigned to the P-site as described. Metagene profiles were assembled by averaging the read densities for one CDS per transcript family, selected as the one found on the transcript with the greatest RNA-seq FPKM value. To be included in the metagene, each CDS was required to include at least 50 aligned reads in all four drug-treatment datasets. Prior to averaging, the read densities for each CDS were normalized by their average value (across all of the treatments and footprint lengths). Metagene profiles were prepared for each read length for each of the four treatments.

Transcript sequences were used to identify all NUG-initiated ORFs. For each ORF, if at least one Harr or LTM read was within one nucleotide of the first base of the initiator codon, profiles of expected read density were constructed from the metagene read densities, explicitly modeling the positions from 3 nt upstream to 150 nt downstream of the start codon (153 nt total) and from 18 nucleotides upstream to the end of the stop codon (21 nt total). For ORFs shorter than 168 nucleotides, the explicitly modeled region was truncated as necessary. For ORFs longer than 168 nucleotides, the remaining codons were filled with 3 repeated values obtained by averaging the read densities of all codons within the annotated CDSs, excluding the first 50 and last 7 codons, which were already used to model the density near the start and stop codon. For each prospective start codon, metagene profiles were also constructed to represent "abortive initiation" events, consisting of only the read density on the initiator codon itself and the three nucleotides upstream of it. For the CHX and ND datasets only, profiles were also constructed for each stop codon, spanning the final codon and the stop codon, to account for the possibility that a given stop codon might have especially slow kinetics

Following construction of the metagene profiles, the P-site mapped reads on each transcript were fit using a non-negative least-squares regression (the nnls function of the scipy.optimize Python library). Fits were performed first for the LTM and Harr datasets independently; coefficient values from these regressions were summed for all ORFs initiating at the same genomic coordinate (including the abortive initiation profile), and a corresponding Wald statistic was calculated. Initiators for which no positive coefficients were obtained were discarded from subsequent steps. Wald statistics were calculated using a homoscedastic error model.

Next, non-negative least-squares regressions were performed on the combination of the CHX and ND datasets. Coefficient values from these regressions were summed for all ORFs initiating at the same genomic coordinate excluding the abortive initiation profile, and a corresponding Wald statistic was calculated. Separately, coefficient values

were summed for all ORFs terminating at the same genomic coordinate, including the stop-only profiles, and a corresponding Wald statistic was calculated. ORFs assigned a regression coefficient value of zero were excluded from further analysis. Wald statistics were calculated for ORFs aggregated by shared start or stop codon, rather than for each individual ORF, because ORFs within those sets are frequently nearly (or actually) linearly dependent, resulting in low confidence in the translation of any specific ORF but high confidence that at least one of the set is translated.

Each genomic stop codon may serve as the terminator for multiple ORFs initiating at different codons on the same or alternative transcript isoforms. To evaluate the relative levels of translation initiating at each start codon within the group of ORFs ending at the same stop codon, we assembled the summed regression coefficients for each of those start codons and divided them by their maximum value.

For the purposes of evaluating translation, six features were collected for each ORF: the three Wald statistics for its genomic start from the Harr, LTM, and CHX/ND regressions; the Wald statistic for its genomic stop from the CHX/ND regression; the relative translation level for its genomic start relative to the other ORFs terminating at the same stop codon; and the actual magnitude of the summed regression coefficients for that stop codon (to enable poorly translated ORFs to be penalized). Using these features, a random forest consisting of 2048 trees was trained on all AUG-initiated ORFs at least 100 codons long, using the ORFs initiating and terminating at annotated start and stop codons as the positive set and all of the other such ORFs as the negative set. Annotated start and stop codons were taken from annotations in the Ensembl and UCSC Known Gene transcript collections. The trees of the random forest were required to have at least 16 training examples in each leaf; this value was selected because it maximized cross-validation accuracy among the set {8, 16, 32, 64, 128}. The random forest achieved 85% six-fold cross validation accuracy on the training set and was next applied to all ORFs regardless of length or start codon. We preferred a random forest over other machine learning classifiers because random forests are simple and involve tuning only one free parameter (the minimum number of training examples in each leaf; the algorithm is relatively insensitive to the number of trees as long as a sufficiently large number is used). Other classifiers, such as support vector machines, demand the imposition of a distance metric, which we had no principled mechanism to select. A random forest, in contrast, is sensitive only to each feature's rank, not its magnitude.

All of the features used for classification by the random forest positively correlate with likelihood of translation; however, the random forest does not guarantee that an ORF with greater values for all features will necessarily receive a higher score. This can lead to overfitting, in which sparsity of the training set results in some ORFs being penalized despite having all higher feature values. To counteract this, we applied a monotonization procedure, based on an equivalence to a network flow problem (Picard, 1976; Spouge et al., 2003), to assign a set of final scores that obeyed the monotonicity constraint with minimum sum-of-squares deviation from the raw random forest scores. Briefly, the feature values are used to construct a partial ordering of the ORFs, interpreted as a directed acyclic graph, with each node corresponding to an ORF and directed links indicating ordered pairs of ORFs for which the successor has feature values all greater than or equal to those of the predecessor. The random forest score of each ORF is associated as the "cost". The transitive reduction of this graph (i.e., the minimal directed

graph preserving reachability) is then recursively partitioned: first, a "source" node is connected to each of the nodes in the graph whose cost is above the average cost, with edge weights set to the excess cost relative to the mean; a "sink" node is connected to the other nodes, with each edge's weight set to the difference between the mean cost and that node's cost; the weights of the internal edges are set to an effectively infinite value; the minimum cut algorithm as implemented in the Python igraph library (Csardi and Nepusz, 2006) is applied to the graph, partitioning it into two subgraphs; and the process is recursively applied to the subgraphs, terminating when one of the partitions contains no ORFs but only the source or sink node. At this point, the score of each ORF within each final partition is reset to the average score of all of the ORFs in that partition. High-confidence translated ORFs were defined as those whose final score exceeded 0.8; the extended confidence set was defined as those with score 0.5–0.8.

The ORF-RATER pipeline was implemented in Python and executed on a server housing 64 cores (64-bit Intel Xeon X7560, 2.27 GHz) and 252 GB of RAM. Algorithms were parallelized by chromosome. The slowest steps were the regressions against the expected read profiles: the Harr regression took 7 hours on 11 cores; the LTM regression took 4 hours on 6 cores; and the CHX/nodrug regression took 13 hours on 8 cores. Including read alignment and all other steps, a typical dataset could be processed over the course of a few days.

Novel CDS metagenes and phasing

Metagenes of novel translation events were compiled in much the same way as for the annotated set. Only those novel ORFs long enough to encompass all of the codons in the metagene plot were included in the metagene. The phasing values were obtained by averaging together the reads in the ND dataset of length 29 and 30 in the appropriate positions of each novel CDS (excluding start and stop codons and the codon preceding them); these read lengths showed the strongest periodicity at canonical CDSs.

Quantification of translation

To quantify expression of the high-confidence CDSs within each transcript family at each CHX or ND timepoint, simplified profiles were constructed for each in which the same three values—the average fraction of reads in each coding frame across all of the codons of canonical CDSs in the dataset in question—were assigned to each codon. For each CDS, windows of three codons at the start and stop codons were excluded from the analysis (the start codon and two surrounding codons, and the stop codon and two preceding codons); reads of all lengths at the remaining codons were aggregated and a regression was performed. In this way, the estimated translation at isolated CDSs is a weighted sum in which reads in the proper reading frame contribute more than those in other frames, and the estimated translation of overlapping CDSs is fractionally assigned based on the read density in non-overlapping regions and based on the frame of the reads in the overlapping region (if the CDSs are in different frames). The number of reads assigned to any given CDS were then normalized by CDS length and sequencing depth to produce reads per kilobase per million (RPKM) expression values for each CDS.

Translational dynamics and clustering

In order to investigate expression changes throughout the LPS stimulation time series, we first median-centered the RPKM expression values of all CDSs for each condition and time point. Because CHX and ND expression levels are well correlated, we averaged the two CHX and single ND replicate for every CDS. These averaged values were used for subsequent analysis. We required that all analyzed CDSs have at least a sum of 10 averaged RPKM across the time series. For each CDS, a simple yet robust translational "fold change" was defined as its maximum translation level divided by its minimum translation level, after smoothing the translation time series using a sliding window of three time points. For example, the fold change assigned to a CDS whose translation monotonically declined was the average of the first three time points divided by the average of the final three time points. The averaging procedure served to buffer against outlying measurements. CDSs whose fold change exceeded a value of 2 were selected for hierarchical clustering. Z scores were calculated for each CDS, and Pearson correlation was used to quantify pairwise similarity (using the linkage function of the scipy.cluster.hierarchy library). Based on these pairwise distances, 128 clusters of CDSs were identified using the fcluster function of the scipy.cluster.hierarchy library.

Gene ontology analysis

From the 128 clusters formed by hierarchical clustering, 3 prominent examples were chosen because they exhibit robust and distinct expression patterns, peaking in expression early, mid, and late in the time series. For each of these three clusters, unique annotated CDSs were identified and searched for enriched Gene Ontology Terms using The Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang et al., 2008). We employed all Biological Process Terms (GOTERM_BP_ALL). To limit redundancy in the final output, we made use of DAVID's functional annotation clustering, which groups GO terms based on similar gene members. We plot the top five GO terms for each of the clusters identified in Figure 4D and their DAVID enrichment scores, which reflect the geometric mean of the p values for all GO terms included in that annotation cluster.

PhyloCSF

For each novel CDS, those codons for which no nucleotides overlapped annotated ORFs were isolated. If at least ten such codons existed, corresponding sequences were identified in a set of ten mammals spanning the Euarchontoglires and whose genome assembly was in at least its second iteration: human, chimpanzee, rhesus macaque, bushbaby, mouse, rat, guinea pig, squirrel, rabbit, and pika. Aligned sequences were retrieved using the 60-way multispecies alignment available from the UCSC genome browser. PhyloCSF software (Lin et al., 2011) was downloaded and applied to the aligned sequences if they could be identified in at least five of the species. The background distribution of PhyloCSF scores for uORFs were obtained by taking the set of all AUG-initiated uORFs of high-confidence translated canonical CDSs (to avoid uORFs on spurious transcripts) that were at least 10 codons long and non-intersecting with any annotated CDSs. Similarly, the background distribution of PhyloCSF scores for extensions were obtained for the set of all NUG-initiated extensions of translated CDSs that spanned at least 10 codons, taking the minimal such extension for each CDS and requiring no overlaps with any other canonical CDSs. The distribution of PhyloCSF

scores at intergenic regions were taken from a random sampling of AUG-initiated intergenic ORFs of at least 10 codons.

Correspondence of BMDC and HFF CDSs

Following application of the ORF-RATER algorithm to both mouse and human ribosome profiling datasets, liftover software (Hinrichs et al., 2006) was used to map the genomic coordinates of the mouse translated initiator codons to the human genome. The corresponding human initiator codon was defined as the nearest NUG up to a maximal distance of 9 nucleotides; of the 10,634 translated mouse initiators, 9827 (92%) corresponding human genomic positions were identified, of which 9056 (92%; 85% overall) could be assigned to an initiator codon on a HFF transcript. Of those, 6186 (68%; 58% overall) had been identified by ORF-RATER as high-confidence translation initiation sites in the HFF dataset. For translation initiation sites potentially leading to the translation of multiple CDSs (on different transcript isoforms), the CDS length plotted in Figure 5A was selected as the one assigned the highest ORF-RATER score. For the gene-level analysis (Figure 5B), translation initiation sites for which a corresponding site could be identified in HFFs were grouped for each family of mouse transcripts, and the total number and the number for which a corresponding translation event in HFFs was identified was calculated.

Identification of peptides and proteins

First, we generated a concatenated database including all canonical proteins in the mouse UniProt database (July 2014), additional annotated proteins from Ensembl and UCSC Known Gene that were missing from the UniProt database, and new proteins identified by ORF-RATER (including scores 0.5 and above). The final search database included 64,997 entries. All mass spectra were analyzed with MaxQuant software version 1.5.2.8 (Cox et al., 2011; Cox and Mann, 2008) as previously described (Jovanovic et al., 2015), and with Spectrum Mill software package v4.0 beta (Agilent Technologies, Santa Clara, CA) as previously described (Mertins et al., 2012; Mertins et al., 2014). For both software packages, we used default parameters and applied a maximum FDR of 1% separately on the protein and peptide level. Allowed variable modifications followed program defaults: methionine oxidation and N-terminal acetylation for both programs, and asparagine deamidation to aspartate for Spectrum Mill. Only peptides uniquely matching to newly identified proteins were considered to be positive identifications of a novel translation event and included in Table S2. For Table S3, each peptide was assigned to the most parsimonious protein from which it could have arisen. To make these assignments, all potential protein assignments were identified for each peptide; next, proteins were selected one by one that could explain the maximum number of unassigned peptides, until all of the peptides had been assigned a source. If more than one protein was available that could explain the same number of peptides, the one with the maximal ORF-RATER score was selected. The MS dataset generated should be considered of qualitative nature. Although it is a mix of BMDCs stimulated for different length of time with LPS, the major goal was to increase proteome coverage, i.e., increase the chance of detecting proteins during any phase of LPS stimulation, and not to provide quantitative differences of protein expression between the time points.

Length distributions

Empirical nucleotide abundances were calculated from the set of mouse BMDC transcripts. Based on these abundances, the frequency of stop codons was calculated to be 1/20. The distribution of ORF lengths on scrambled transcripts (Figure 3A) was calculated as a geometric distribution with that parameter. The length distribution of previously annotated CDSs was plotted based on the CCDS collection (Farrell et al., 2014).

Protein multiple sequence alignments

The codon sequences of the proteins plotted in Figures S3B and S6A were obtained for each mammal using the 60-way multispecies alignment, followed by translation to amino acid sequence and alignment using Clustal omega (Sievers et al., 2011). Alignments were plotted using Jalview (Waterhouse et al., 2009).

Cloning and cell culture

To generate the plasmid constructs used in Figure S4A, the first exon of the human *Fxr2* gene was amplified from purified human genomic DNA (courtesy M. Leonetti) by PCR using primers 5′-CGA TTG ACT GAG TCG CCC GGA TCC GCA GTA GGC GGC GGT G-3′ (forward) and 5′-CTC CTC GCC CTT GCT CAC ACC AGA ACC ACC CTT GTA GAA GGC CCC GTT GG-3′ (reverse), where the underlined segment is the portion complementary to the genomic DNA sequence. Similarly, for the constructs of Figure S4B, the 5′ portion of human *Zdhhc3* was amplified from purified human cDNA (courtesy C. Jan) using primers 5′-CGA TTG ACT GAG TCG CCC GGA TCC GCG TCA TCA ACC TGC GCG G-3′ (forward) and 5′-CTC CTC GCC CTT GCT CAC ACC AGA ACC ACC ACA GGC GAT GCC ACA GCC-3′ (reverse). Amplified fragments were purified by polyacrylamide gel electrophoresis, cloned using the Zero Blunt TOPO kit (ThermoFisher), and sequenced. Specific point mutants (indicated in Figure S4) were generated by PCR using primers harboring appropriate mismatches. Cloned fragments were then amplified by PCR and fused with an eGFP sequence and inserted into the LeGO-iC2 vector (Weber et al., 2008) using HiFi assembly (New England Biolabs). The CDS of human *MMP24-AS1* was amplified from human genomic DNA using PCR primers 5′-CGA TTG ACT GAG TCG CCC GGA TCC GAG ACC ATG GGG GCT CAG CTA AG-3′ (forward) and 5′-GCT CCT CGC CCT TGC TCA CAC CGG AGC CAC CGG TGT ACC AGG AAG TGC AGG CGA TG-3′ (reverse) and inserted into the LeGO-G2 vector using HiFi assembly. Inserts in all final constructs were validated by sequencing.

For fluorescence measurements and microscopy, HEK 293T or HeLa cells were grown in DMEM medium with high glucose, supplemented with glutamine, penicillin/streptomycin, and 10% FBS. Constructs were transfected using TransIT-LT1 (Mirus) two days prior to data collection. For the *Fxr2* and *Zdhhc3* experiments (Figure S4), constructs were transfected into 293T cells, and eGFP and mCherry fluorescence were quantified using a BD Bioscience LSR-II flow cytometer. Imaging of eGFP-tagged *MMP24-AS1* was performed in HeLa cells co-transfected with ER-mCherry or Golgi-mCherry constructs as indicated, cultured on a 24-well #1.5H glass bottom plate (Cellvis). Cells were imaged live on a widefield epifluorescence microscope using a

Lambda LS illuminator (Sutter), 100X Nikon Plan Apo VC 1.4 oil objective, and Andor Clara CCD camera, controlled with Micro-Manager software (Edelstein et al., 2010).

SUPPLEMENTAL REFERENCES

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26,* 1367-1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. *10,* 1794-1805.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems *1695,* 1-9.

Edelstein, A., Amodaj, N., Hoover, K., Vale, R., and Stuurman, N. (2010). Computer control of microscopes using μManager. Current Protocols in Molecular Biology 14.20. 1-14.20. 17.

Farrell, C.M., O'Leary, N.A., Harte, R.A., Loveland, J.E., Wilming, L.G., Wallin, C., Diekhans, M., Barrell, D., Searle, S.M., Aken, B., *et al.* (2014). Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Res. *42,* D865-72.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K.,Jr, Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., and White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31,* 5654-5666.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., *et al.* (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34,* D590-8.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4,* 44-57.

Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., *et al.* (2015). Dynamic profiling of the protein life cycle in response to pathogens. Science *347,* 1259038.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27,* i275-82.

Mertins, P., Qiao, J.W., Patel, J., Udeshi, N.D., Clauser, K.R., Mani, D., Burgess, M.W., Gillette, M.A., Jaffe, J.D., and Carr, S.A. (2013). Integrated proteomic analysis of post-translational modifications by serial enrichment. Nat. Methods *10,* 634-637.

Mertins, P., Udeshi, N.D., Clauser, K.R., Mani, D.R., Patel, J., Ong, S.E., Jaffe, J.D., and Carr, S.A. (2012). iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. Mol. Cell. Proteomics *11*, M111.014423.

Mertins, P., Yang, F., Liu, T., Mani, D.R., Petyuk, V.A., Gillette, M.A., Clauser, K.R., Qiao, J.W., Gritsenko, M.A., Moore, R.J., *et al.* (2014). Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. Mol. Cell. Proteomics *13*, 1690-1704.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X.J., Harte, R., Balasubramanian, S., Tanzer, A., and Diekhans, M. (2012). The GENCODE pseudogene resource. Genome Biol. *13*, R51.

Picard, J. (1976). Maximal closure of a graph and applications to combinatorial problems. Manage. Sci. *22*, 1268-1272.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protoc. *2*, 1896-1906.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., and Lu, D. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236-240.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., and Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Spouge, J., Wan, H., and Wilbur, W. (2003). Least squares isotonic regression in two dimensions. J. Optim. Theory Appl. *117*, 585-605.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T., Hein, M.Y., Huang, S.X., Ma, M., Shen, B., Qian, S.B., Hengel, H., *et al.* (2012). Decoding human cytomegalovirus. Science *338*, 1088-1093.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Weber, K., Bartsch, U., Stocking, C., and Fehse, B. (2008). A multicolor panel of novel lentiviral “gene ontology” (LeGO) vectors for functional gene analysis. Molecular Therapy *16*, 698-706.