

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

A method for annotating protein-coding sequences from ribosome profiling data

Permalink

<https://escholarship.org/uc/item/79j0n8qd>

Author

Rodriguez, Edwin

Publication Date

2014

Peer reviewed|Thesis/dissertation

A method for annotating protein-coding sequences
from ribosome profiling data

by

Edwin Hadalid Rodriguez

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

Copyright 2014
by
Edwin Hadalid Rodriguez

Dedication

To my family and friends.

This work would not have been possible without the help of Alexander P. Fields, Marko Jovanovic, and Brian Haas.

A method for annotating protein-coding sequences from ribosome profiling data

Edwin Hadalid Rodriguez

Abstract

Coding sequences (CDSs) have traditionally been predicted on the basis of sequence and length; on any given mRNA the longest ATG-initiated CDS is annotated as the translated protein product. Ribosome profiling, a technique for sequencing regions of mRNA being decoded by the ribosome, enables a high-resolution view of translation and is thus a method for the data-dependent annotation of CDSs. When performed in the presence or absence of drugs that target the ribosome in stereotyped ways, ribosome profiling reveals the core features of translation: initiation, phased elongation, and termination.

We developed and applied a computational analysis pipeline that leverages all these features to identify CDSs systematically. Notably, our analysis is not biased by length, initiator codon, or overlapping CDS structure. We performed our experiments in bone marrow-derived dendritic cells undergoing lipopolysaccharide (LPS) stimulation. Cell type-specific transcript assemblies guided our analysis, and mass spectrometry corroborated our predictions. We identified translation of both annotated and unannotated CDSs.

Unannotated translation comprises previously unknown variants of annotated proteins and completely novel CDSs. We identified 829 N-terminal truncations and 609 N-terminal extensions, many of which affect localization signals or domain structure. We identified 5,215 CDSs translated from regions upstream of annotated CDSs; these upstream CDSs may regulate expression of their associated downstream CDS. We identified 472 new CDSs that arise from new transcripts or transcripts previously defined as non-coding. Finally, we found 989 translation events that occur within but out-of-frame relative to annotated CDSs.

We used ribosome density to quantify expression of all identified CDSs at various times after LPS stimulation (0, 0.5, 1, 2, 4, 6, 8, 9, 12 hrs). Many newly identified CDSs display significantly enhanced or repressed expression through the time-course, relative to mock-stimulated cells.

Table of contents

Preface	Dedication	iii
	Abstract	iv
	Table of contents	vi
	List of figures	viii
Chapter 1	Introduction	1
	A history of coding sequence annotations	2
	Recent efforts to annotate protein-coding sequences	4
	Figures	7
	References	8
Chapter 2	Methods for annotating coding sequences	10
	Introduction	11
	Dendritic cell isolation	11
	Collection of LPS-stimulated BMDCs for ribosome profiling	12
	Processing of samples for ribosome profiling	13
	Alignment of sequencing reads	16
	P-site mapping of aligned reads	16
	Creating an idealized CDS profile	17
	A regression-based approach for CDS annotation	17
	DC protein isolation and processing for subsequent mass spectrometry	19
	LC-MS/MS measurements	20

	MS identification and quantification of proteins	21
	Figures	22
	References	24
Chapter 3	Results	26
	Results	27
	Figures	31
Chapter 4	Discussion	41
	Discussion	42

List of figures

Chapter 1		
Figure 1-1	Short proteins are underrepresented across the tree of life	7
Chapter 2		
Figure 2-1	Features of translation observed in ribosome profiling data	22
Figure 2-2	A final classification score captures known proteins and uncovers novel translation events	23
Chapter 3		
Figure 3-1	A CDS annotation pipeline and summary findings	32
Figure 3-2	Truncated ORFs	33
Figure 3-3	Extended ORFs	34
Figure 3-4	New ORFs	35
Figure 3-5	Upstream ORFs	36
Figure 3-6	Start-overlap ORFs	37
Figure 3-7	Downstream ORFs	38
Figure 3-8	Internal ORFs	39
Figure 3-9	Expression changes in newly annotated ORFs upon LPS stimulation of DCs	40

Chapter 1

Introduction

A history of coding sequence annotation

The sequencing of the first eukaryotic chromosomes in the 1990s made it possible to enumerate the number of proteins they encode (Dujon et al., 1994; Goffeau et al., 1996; Oliver et al., 1992). In *S. cerevisiae*, where introns are rare, one could in principle simply count the number of non-overlapping open-reading frames (ORFs), *i.e.*, ATG codons followed by in-frame stop codons, that appeared in the DNA sequence to arrive at the number of possible protein-coding regions. Yet such a simple annotation algorithm yielded a number of ORFs too vast to be believable (see (Basrai et al., 1997)). To constrain the ORF prediction algorithm, a null sequence model of the same size and composition as yeast chromosome 11 was generated and searched for ORFs. Such a random sequence yielded mainly ORFs shorter than 100 codons, and all ORFs identified were less than 150 codons long (Dujon et al., 1994). These results suggested that ORFs longer than 100 codons were likely to be protein-coding, especially when codon usage mirrored that of known genes, and that proteins shorter than 100 amino acids could not be confidently annotated from DNA sequence alone. Thus emerged three biases in coding sequences (CDS) annotation pipelines that have been propagated since from fungi to mammals (Dujon et al., 1994; Okazaki et al., 2002): in general, CDSs have only been annotated when they initiate with an ATG codon, are not contained within other CDSs, and are longer than 100 codons.

It was clear that these simple rules, while well reasoned and useful on a genome-wide scale, were violated in many biological contexts, and thus systematically excluded true translation events. Non-canonical translation initiation at a CTG codon had been demonstrated for the human transcription factor c-myc both as a possibility *in vitro* but also as a feature of tumor-derived human cell lines (Hann and Eisenman, 1984; Hann et al., 1988). Remarkably, translation of the human growth factor FGF2 was shown to begin from multiple CTG codons,

and one of these CTG-initiated forms is the primary expressed variant of FGF2 in human placenta (Prats et al., 1989; Sommer et al., 1987). Viral and prokaryotic genomes were known to encode many overlapping out-of-frame CDSs (Normark et al., 1983), a finding that has recently been extended to mammalian genomes (Bergeron et al., 2013; Poulin et al., 2003; Vattem and Wek, 2004).

In yeast, where these annotation standards were established, the fact that many core biological processes make use of short proteins suggested that the 100 amino acid cut-off was indeed arbitrary. In particular, it was known that multiple ribosomal proteins are 25-59 amino acids long; several components of the ATP synthase complex range in size from 48-76 amino acids; TOM5 and TOM6, involved in protein import to the mitochondria, are 50 and 61 amino acids; OST4, a transmembrane protein important for assembly of the oligosaccharyltransferase complex is only 36 amino acids long; and full length MFA1 and MFA2, the mating pheromones of yeast, are 36 and 38 amino acids (Basrai et al., 1997).

Furthermore, early mRNA sequencing techniques revealed that many yeast transcripts encoding short ORFs were regulated in a cell cycle-dependent manner, suggesting that their short protein products were functional. Indeed, functional genomics screens provided evidence that these unannotated short proteins were critical for growth under conditions of high temperature, DNA damage, and a non-fermentable carbon source (Basrai et al., 1999; Kastenmayer et al., 2006). Intriguingly, when the distribution of sizes for the set of high-confidence protein annotations (UniProt Swiss-Prot set) is plotted for various distantly related organisms, there is a conspicuous dearth of proteins shorter than 100 amino acids (Fig. 1-1), suggesting that many functional short proteins remain to be discovered across the tree of life.

Recent efforts to annotate protein-coding sequences

The works described thus far were all attempts to make genome-wide assertions about translation events from DNA or RNA sequence data without the use of a tool that actually captured information about translation. A technique that captures peptide-level evidence of translation on this scale is mass spectrometry (MS). Recent advances in instrumentation have dramatically increased the sensitivity of this technique and offered the promise of annotating all proteins encoded by the genome (Hu et al., 2005; Kim et al., 2014; Wilhelm et al., 2015). The details of the MS workflow, however, introduce their own biases that have often reinforced those of the original genome annotations. Generally, clarified whole-cell lysates are treated with a protease, typically trypsin, resulting in peptides with charged terminal amino acid side chains that increase the ability of the instrumentation to detect the peptide and aid in assignment of the peptide to a protein sequence. The masses of these peptides are measured and a database of possible tryptic peptides derived from *in silico* digested full protein sequences is searched for matches. Two biases arise from the workflow described: proteins that are not effectively digested because of amino acid composition are not likely to be detected, and it is generally not possible to discover novel translation events because one must start with a database of predicted proteins. The MS field has implemented a solution for each of these problems by using multiple proteases, *e.g.*, LysC in addition to trypsin, and by populating MS search databases with all possible ORFs found on a given set of mRNAs. Even with these corrections, an amino acid composition bias remains in what can be discovered, and the size of the search database must be restricted to maintain reasonable false discover rates by considering only ATG-initiated ORF, setting a length cut-off on ORFs considered, or restricting the number of RNAs on which one searches for ORFs.

Thus, standard mass spectrometry is best thought of as a tool for confirming predicted CDSs than for CDS annotation.

Ribosome profiling is a recently developed technique that allows one to sequence regions of mRNA occupied by translating 80S ribosomes, so-called ribosome footprints (Ingolia et al., 2009). When aligned to the genome or transcriptome, the vast majority of these footprints fall within regions of annotated CDSs. These footprints are 28-34 base pairs in length, which corresponds to the length of mRNA covered by a translating ribosome, and the footprint size distribution can be shifted by translation elongation inhibitors known to alter ribosome structure (Ingolia et al., 2014). Importantly, these footprints reveal core features of translation: they are strongly enriched at annotated starts and stops of CDSs and display clear 3-base pair periodicity (Ingolia et al., 2011). Furthermore, these features can be exaggerated when ribosome profiling is coupled with drugs that inhibit the ribosome in different phases of translation (Ingolia et al., 2011; Lee et al., 2012). These features can be appreciated regardless of overlapping CDS structure, initiator codon, and CDS length. Thus, ribosome profiling is the first technique available for the unbiased, data-dependent annotation of CDSs.

We and others have previously used these features independently to find novel translated products in a variety of biological contexts. The mere presence of ribosomes within ORFs was used to annotate upstream ORFs (uORFs) and novel short CDSs in yeast undergoing amino acid starvation and meiosis (Brar et al., 2012; Ingolia et al., 2009). In human fibroblasts infected with cytomegalovirus (CMV), treatment with Harringtonine and Lactimidomycin, translation initiation inhibitors, demonstrated translation initiation sites and allowed for the prediction of novel CMV CDSs (Stern-Ginossar et al., 2012). A similar approach in mouse embryonic stem cells revealed novel CDSs, including some translated from RNAs previously believed to be non-

coding (Ingolia et al., 2011). In Zebrafish, the phased elongation observable in ribosome profiling data has been used to annotate translated ORFs (Bazzini et al., 2014). Finally, the drop in ribosome footprint counts that occurs in the transition from coding sequence to 3' untranslated region (UTR) has been used as a metric to evaluate the coding potential of RNAs (Guttman et al., 2013).

Figures

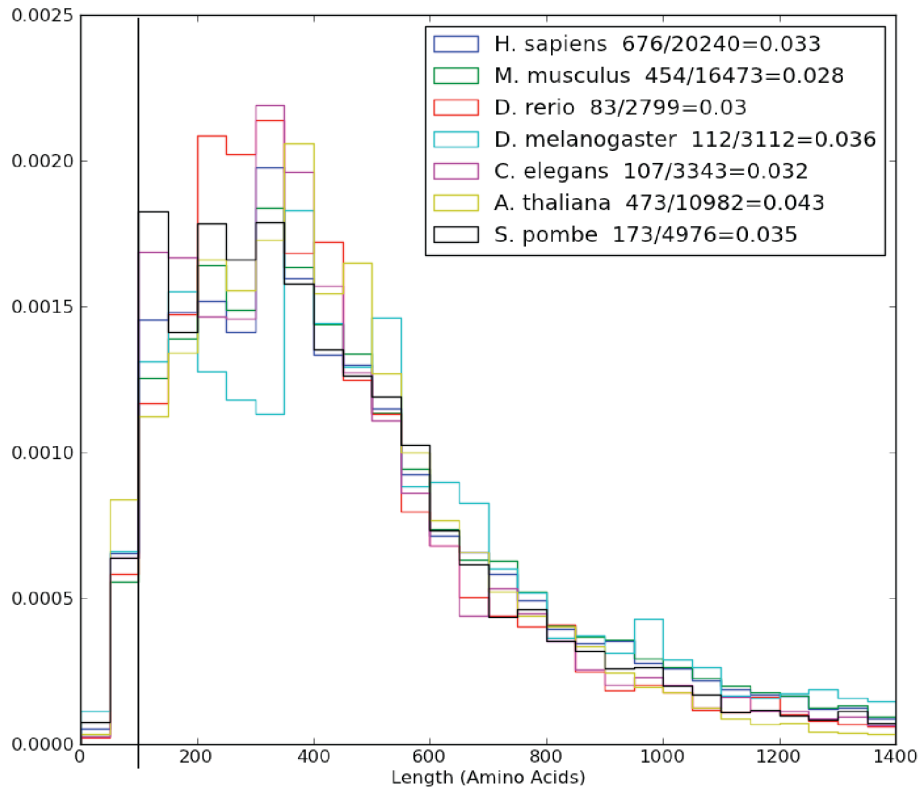


Figure 1-1. Short proteins are underrepresented across the tree of life. Normalized histogram for the distribution of protein sizes for each noted organism. In the legend appears the fraction of annotated proteins under 100 amino acids in each organism. The protein sets used for the graphic are the manually-curated UniProt Swiss-Prot sets.

References

- Basrai, M.A., Hieter, P., and Boeke, J.D. (1997). Small open reading frames: beautiful needles in the haystack. *Genome Res.*
- Basrai, M.A., Velculescu, V.E., Kinzler, K.W., and Hieter, P. (1999). NORF5/HUG1 is a component of the MEC1-mediated checkpoint response to DNA damage and replication arrest in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* *19*, 7041–7049.
- Bazzini, A.A., Johnstone, T.G., and Christiano, R. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO*
- Bergeron, D., Lapointe, C., Bissonnette, C., Tremblay, G., Motard, J., and Roucou, X. (2013). An Out-of-frame Overlapping Reading Frame in the Ataxin-1 Coding Sequence Encodes a Novel Ataxin-1 Interacting Protein. *Journal of Biological Chemistry* *288*, 21824–21835.
- Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* *335*, 552–557.
- Dujon, B., Alexandraki, D., André, B., Ansorge, W., Baladron, V., Ballesta, J.P., Banrevi, A., Bolle, P.A., Bolotin-Fukuhara, M., and Bossier, P. (1994). Complete DNA sequence of yeast chromosome XI. *Nature* *369*, 371–378.
- Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. *Science* *274*, 546–563–7.
- Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* *154*, 240–251.
- Hann, S.R., and Eisenman, R.N. (1984). Proteins encoded by the human c-myc oncogene: differential expression in neoplastic cells. *Mol. Cell. Biol.*
- Hann, S.R., King, M.W., Bentley, D.L., and Anderson, C.W. (1988). A non-AUG translational initiation in c- myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. *Cell* *52*, 185–195.
- Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005). The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* *40*, 430–443.
- Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *CellReports* *8*, 1365–1379.
- Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* *324*,

218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* 147, 789–802.

Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.-C., Yang, H., Carter, C.D., Wheeler, D., Davis, R.W., Boeke, J.D., et al. (2006). Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res.* 16, 365–373.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. *Nature* 509, 575–581.

Lee, S., Liu, B., and Huang, S.X. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution.

Normark, S., Bergström, S., Edlund, T., Grundström, T., Jaurin, B., Lindberg, F.P., and Olsson, O. (1983). Overlapping genes. *Annu. Rev. Genet.* 17, 499–525.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573.

Oliver, S.G., Van der Aart, Q., and Agostoni-Carbone, M.L. (1992). The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46.

Poulin, F., Brueschke, A., and Sonenberg, N. (2003). Gene Fusion and Overlapping Reading Frames in the Mammalian Genes for 4E-BP3 and MASK. *Journal of Biological Chemistry* 278, 52290–52297.

Prats, H., Kaghad, M., Prats, A.C., Klagsbrun, M., Lélias, J.M., Liauzun, P., Chalon, P., Tauber, J.P., Amalric, F., and Smith, J.A. (1989). High molecular mass forms of basic fibroblast growth factor are initiated by alternative CUG codons. *Pnas* 86, 1836–1840.

Sommer, A., Brewer, M.T., Thompson, R.C., Moscatelli, D., Presta, M., and Rifkin, D.B. (1987). A form of human basic fibroblast growth factor with an extended amino terminus. *Biochem. Biophys. Res. Commun.* 144, 543–550.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K., 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.

Vattem, K.M., and Wek, R.C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11269–11274.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2015). Mass-spectrometry-based draft of the human proteome. *Nature* 509, 582–587.

Chapter 2

Methods for annotating coding sequences

Introduction

Here we describe the ribosome profiling-based CDS annotation of mouse bone marrow-derived dendritic cells (BMDCs) undergoing lipopolysaccharide (LPS) stimulation. The regression-based algorithm described herein is used to annotate CDSs on a canonical set of transcripts from UCSC and ensembl and on a set of BMDC-specific transcripts. MS was used to corroborate expression of the CDSs that the new algorithm annotates.

Our method is the first to leverage all the features of translation that can be appreciated from ribosome profiling data: initiation peaks, phased elongation, and termination peaks. These features become readily observable when one creates a metagene profile of all annotated protein-coding sequences (Fig. 2-1). The translation initiation drugs Harringtonine (Harr) and Lactimidomycin (LTM) aid in increasing ribosome density at initiation sites; Cycloheximide (CHX)-treated and untreated samples display phasing within the body of the ORF; and the untreated samples accumulate ribosomes at translation termination sites.

Dendritic cell isolation

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 modified version of the DC isolation protocol as described previously (Amit et al., 2009; Chevrier et al., 2011; Garber et al., 2012; Lutz et al., 1999; Rabani et al., 2011). 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 (50uM, Invitrogen), L-glutamine (2mM, VWR), penicillin/streptomycin (100U/ml, VWR), MEM non-essential amino acids (1X, VWR), HEPES

(10mM, VWR), sodium pyruvate (1mM, 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), 100mm non tissue culture treated plastic dishes using 10ml medium per plate. At day 2, cells were fed with another 10ml medium per dish. At day 5, cells were harvested from 15ml of the supernatant by spinning at 1,400 rpm for 5 minutes; pellets were resuspended with 5ml medium and added back to the original dish. Cells were fed with another 5ml medium at day 7. At day 8, all nonadherent and loosely bound cells were collected and harvested by centrifugation. Cells were then resuspended with medium, plated at a concentration of 15×10^6 cells in 10ml medium per 100mm dish. At day 9, cells were stimulated for various time points with LPS (100ng/ml, rough, ultrapure E. coli K12 strain, Invitrogen).

Collection of LPS-stimulated BMDCs for ribosome profiling

BMDCs were stimulated with LPS for 0, 0.5, 1, 2, 4, 6, 8, 9, and 12 hours. A control set of cells was mock stimulated with DMSO alone.

To annotate translation initiation sites across the stimulation time course, cells from the indicated time points were collected and pooled. Pooled cells were treated with the translation initiation inhibitors Harringtonine (Harr, 1ug/mL final concentration, LKT Laboratories) or Lactimidomycin (LTM, 50ug/mL final concentration, Millipore). In the case of Harr treatment, cells were incubated with the drug for 5 minutes at 37°C, after which Cycloheximide (CHX, 100ug/mL final concentration, Sigma-Aldrich) was added to the culture and incubated for 1 minute at 37°C. In the case of LTM treatment, cells were incubated with the drug for 30 minutes at 37°C. To observe translation elongation and termination, cells from the indicated time points were either treated with the elongation inhibitor CHX (100ug/mL final concentration) for 1

minute at 37°C or left untreated (hereafter referred to as the no-drug sample).

After the indicated treatments, media from the cell culture plates was removed and the plates were placed on ice. Drug-treated samples were washed twice with 5mL of ice-cold phosphate-buffered saline (PBS) supplemented with 100ug/mL CHX, and no-drug samples were washed twice with 5mL of ice-cold PBS alone. Cells were then covered with 400uL of lysis buffer (4.75 mL polysome buffer + 250 uL 20% Triton X-100 (Sigma-Aldrich) + 60 units Turbo DNase (Ambion)). Polysome buffer is 20mM Tris-HCl pH 7.56, 150mM NaCl, 5mM MgCl₂, 100ug/mL CHX, 1mM DTT, and 8% glycerol. Lysis was carried out in the cell culture dish by agitating the dish, scraping the cells off the surface, and triturating 10 times with lysis buffer. Whole-cell lysate was then clarified by centrifuging at 20,000 xg for 10 minutes at 4°C. Clarified lysate was collected and flash frozen in liquid nitrogen.

Processing of samples for ribosome profiling

Total RNA for each sample was quantified using Quant-iT RiboGreen RNA Assay Kit (Life Technologies). Lysates were then treated at room temperature for 45 minutes with 2.4 units of RNase I (Ambion) per ug of total RNA. This treatment results in the digestion of RNA regions not protected by the ribosome, RNA-binding proteins, or RNA secondary structure. The digestion reaction was quenched with 400 units of SUPERase-In (Ambion) and placed on ice. Digested cell lysates were then layered over 1mL of a 34% sucrose cushion and centrifuged for 4 hours at 70,000 RPM at 4°C in a Beckman-Coulter TLA-110 rotor to isolate ribosomes and the ribosome-protected RNA footprints. The sucrose solution was aspirated, and the ribosome pellet resuspended in 500ul of TRIzol (Life Technologies). 100ul of chloroform were added and the solution incubated at room temperature for 3 minutes. The aqueous and non-aqueous phases

were separated by centrifuging for 15 minutes at 4°C at 12,000 xg. The aqueous phase was collected and RNA from it precipitated by adding 2ul of Glycoblue coprecipitant (Life Technologies) and 500uL ice-cold isopropanol. The precipitation solution was vortexed then incubated at room temperature for 10 minutes. RNA was pelleted by centrifuging for 20 minutes at 4°C at 20,000 xg. Supernatant was removed and the RNA pellet washed with 1mL 75% ice-cold ethanol. The RNA was air-dried and resuspended in 25ul of 10mM Tris pH 7. 5ug of sample was treated with the Ribo-Zero rRNA removal kit (Epicentre), following the instructions of the manufacturer. rRNA-subtracted sample was then size-separated by running through a 15% TBE-Urea gel (Life Technologies) in 1X RNase-free TBE (Ambion). Gels were stained with SYBR Gold, and regions corresponding to 28-34 base pairs were cut from the gels. These gel slices were macerated and incubated in 500ul of diethylpyrocarbonate (DEPC)-treated water (Ambion) at 70°C for 10 minutes. Extracted RNA was separated from the gel by centrifuging the gel slurry for 3 minutes at 20,000 xg over a Costar Spin-X column (Corning). RNA was precipitated from the recovered eluate by adding 62.75ul of 3M NaOAc pH 5.5, 2ul of Glycoblue, and 750ul of ice-cold isopropanol and incubating the mix at -30°C for 30 minutes. RNA was pelleted by centrifuging for 30 minutes at 4°C at 20,000 xg. RNA pellet was washed with 750ul of 80% ice cold ethanol and resuspended in 25ul of 10mM Tris, pH 7. Size-selected RNA was dephosphorylated with 250 units of T4 polynucleotide kinase (New England Biolabs) at 37°C for 1 hour. RNA was then precipitated by adding 150ul of DEPC water, 25ul of 3M NaOAc pH 5.5, 2ul glycoblue, and 750uL isopropanol and incubating at -30°C for 30 minutes. RNA was pelleted by centrifuging for 30 minutes at 20,000 xg at 4°C. RNA pellets were washed with 750 uL of 80% ice-cold ethanol and resuspended in 8ul of 10mM Tris pH 7. The 3' ends of these dephosphorylated RNA products were ligated to a 5'-adenylated and 3'-dideoxy-blocked

oligonucleotide (/5rApp/CTGTAGGCACCATCAAT/3ddC/, hereafter referred to as linker-1) by adding 4ul of dephosphorylated RNAs to 8ul 50% PEG8000, 2uL DMSO, 2uL 10X RNA ligase buffer, 1uL SUPERase-In, 1uL linker-1, 2uL T4 RNA ligase 2, and incubating the reaction at room temperature for 2 hours. RNA was precipitated in a manner described above, resuspended in 6ul 10mM Tris pH 7, and size-separated by running on a 10% TBE-Urea gel in 1X RNase-free TBE at 200V for 45 minutes. Ligated products were cut from the gel and purified as described above and resuspended in 20ul 10mM Tris pH 7. 10uL of ligated RNA were reverse-transcribed with the following oligonucleotide 5'-

pGATCGTCGGACTGTAGAACTCTGAACCTGTCG/(18-atom hexa-ethyleneglycolspacer)/CAAGCAGAAGACGGCATAACGAGATATTGATGGTGCCTACAG-3', hereafter referred to as oCJ200. The reverse transcription reaction included the following: 10uL linker-1-ligated RNA, 3.28uL 5X First Strand Buffer, 0.82uL 10mM dNTPs, 0.5uL 100uM oCJ200, 0.5uL SUPERase-In, 0.82uL 100mM DTT, and 0.82uL Superscript III (Life Technologies). This reaction mix was incubated at 50°C for 30 minutes. RNA was non-specifically degraded by adding 1.8 uL of 1M NaOH and incubating at 98°C for 20 minutes. This solution was then size-separated by running on a 10% TB-Urea gel at 200V for 65 minutes. Reverse transcribed products were gel-extracted as described above, precipitated, and resuspended in 15ul of 10mM Tris pH 8. These products were circularized with CircLigase ssDNA Ligase (Epicentre) by adding to following to 15uL of reverse transcribed products: 2uL 10X CircLigase Buffer, 1uL 1mM ATP, 1uL 50mM MnCl₂, and 1ul CircLigase. These reactions were incubated at 60°C for 1 hour, after which the enzyme was heat-inactivated by incubating at 80°C for 10 minutes. Circularized products were amplified via the polymerase chain reaction using the Phusion High Fidelity PCR kit (Thermo Scientific) and the following primers: 5'

CAAGCAGAAGACGGCATAACGA 3' and a barcoded primer 5'

AATGATACGGCGACCACCGAGATCTACACGATCGGAAGAGCACACGTCTGAACTCC
AGTCAC[XXXXXX]CGACAGGTTTCAGAGTTC 3', where XXXXXX represents the hexamer
bar code sequence. These PCR products were gel-purified as described above and sequenced on
the Illumina HiSeq2500 at the UCSF Center for Advanced Technology.

Alignment of sequencing reads

Sequencing reads were stripped of linker-1 sequence using `fastx_clipper` from the FASTX-
Toolkit with the following settings: `-a CTGTAGGCACCATCAAT -n -Q 33 -i [input filename] -`
`o [output filename]`. Reads were then filtered using `bowtie2` (Langmead and Salzberg, 2012) to
remove any reads that mapped to rRNA or mitochondrial rRNA using the following settings `--`
`local -x [bowtie2 index of rRNA and mitochondrial rRNA sequences] -U [input filename] -S`
`[output filename]`. Remaining reads were aligned using `Tophat2` (Kim et al., 2013) to the union
of mm9 genome build transcripts from the UCSC known canonical gene set, `ensembl`, and a
previously published DC-specific transcript set (Shalek et al., 2013). `Tophat2` was run with the
following settings `--b2-very-sensitive --transcriptome-only --no-novel-juncs -max-multihits=20 -`
`-output-dir [output directory name] --transcriptome-index [bowtie2 index of union of transcript`
`models] [bowtie2 index of mm9 genome] [input filename]`.

P-site mapping of aligned reads

These transcript-aligned ribosome footprint reads were then processed so that each read
was only counted at the position that corresponds to the P-site position of the ribosome. In order
to do this, we took advantage of the empirical observation that in all of our datasets, there is an

increased density of reads at the start codon of annotated CDSs that represent initiating ribosomes with their P-sites aligned with the initiation codon. We collected all reads that overlapped annotated CDS start sites; these reads ranged in size from 28-34 base pairs. For reads of any one particular length, we asked how far the 5' end of the read was from the annotated start codon; though a distribution of distances exists, there is a peak in the distance distribution because the majority of ribosome P-sites in this subset of reads are indeed directly over the initiation codon. This peak in the distance distribution was then used to adjust the alignment of the read, so that instead of assigning the read to the whole sequence to which it corresponds, we could assign it to a single nucleotide, the first nucleotide of the initiation codon. This method was used to define the distance by which reads of different sizes ought to be adjusted in order to yield P-site mapped reads, and all transcript-mapped reads (not just those overlapping the initiation codon) were adjusted accordingly.

Creating an idealized CDS profile

In order to annotate CDSs, we first had to construct an idealized profile of a CDS for each of the treatment conditions (Harr, LTM, CHX, and no-drug). We did this by collecting P-site mapped reads from 3 nucleotides upstream to 150 nucleotides downstream of annotated translation start sites and from 21 nucleotides upstream of annotated translation stop sites to the end of the stop codon. These regions were averaged across all annotated genes to create a metagene profile.

A regression-based approach for CDS annotation

We considered all possible NUG-initiated ORFs in our transcript set; there were 7.4 million possible ORFs. We removed ORFs that did not have at least one read within one nucleotide of the start site in either the Harr or LTM dataset; this filter removed 88% of the initially considered ORFs and resulted in a set of 1.6 million ORFs that were possibly being translated. For this set of ORFs, in each treatment condition, the possibility of translation was evaluated via regression analysis and the regression problem posed (described below) was solved using the non-negative least squares approach. Briefly, for any possible ORF, we represented the counts along this region as a vector Y . The expected counts were represented as a matrix X of equal length as Y , and X was populated with values from the idealized profile described above. $Y = mX + b$ was then solved and values of m , the regression coefficients, served as a proxy for the confidence that the models represented in X explain the values observed in Y . Regression coefficients were thus collected for each considered ORF in each of the datasets. Only 302,843 ORFs had any regression coefficient greater than zero. For these ORFs, a random forest classifier was used to classify them as true or false translation events based on the values of their regression coefficients. The random forest classifier was trained on a “gold” set of examples that included all annotated ATG-initiated ORFs greater than 100 codons in length. The random forest classifier combines the multiple regression coefficients into a single score for each ORF; this final score reflects our confidence that any given ORF is being translated. We positively classify 19,511 ORFs with a random forest probability greater than 0.8; this threshold captures most annotated proteins (Fig. 2-2). Importantly, this method of annotating CDSs considers all possible NTG initiator codons, considers ORFs of arbitrary length, and allows for partially or fully overlapping ORFs.

DC protein isolation and processing for subsequent mass spectrometry

After stimulation (LPS or MOCK) and the appropriate time points, 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 μ g/mL aprotinin (Sigma, A6103), 10 μ g/mL leupeptin (Roche, #11017101001), 1 mM PMSF (Sigma, 78830)). Lysates were centrifuged at 20,000g for 10 min, and protein concentrations of the clarified lysates were measured via BCA assay (Pierce). From this procedure, DC lysates produced \sim 100 μ g of protein per 1 million cells. Light- standard and heavy/medium-labeled sample lysates were then combined in a 1:1 protein ratio (20 μ g each and therefore 40 μ g total). Protein disulfide bonds of the combined lysates were reduced for 45 min with 5 mM dithiothreitol (Thermo Scientific, 20291) 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. Peptide mixtures were acidified to a final volumetric concentration of 1% formic acid (Fluka, 56302) and centrifuged at 10,000g for 5 min to pellet urea that had precipitated out of solution. The peptide mixtures were fractionated by Strong Cation Exchange (SCX) using StageTips as previously described (Rappsilber et al., 2007) with slight modifications. Briefly, one StageTip was prepared per sample by 3 SCX discs (3M, #2251) topped with 2 C18 discs (3M, #2215). The packed StageTips were first washed with 100 μ l methanol and then with 100 μ l 80% acetonitrile and 0.5% acetic acid. Afterwards they were equilibrated by 100 μ l 0.5% acetic acid and the sample was loaded onto the discs. The sample was transeparated from the C18 discs to the SCX discs by applying 100 μ l 80% acetonitrile; 0.5% acetic acid, which was followed by 6 stepwise elutions and collections of the peptide mix from the SCX discs. The first fraction was eluted with

50 μ l 50mM NH₄AcO; 20% MeCN (pH 4.1, adjusted with acetic acid), the second with 50 μ l 50mM NH₄AcO; 20% MeCN (pH 4.8, adjusted with acetic acid), the third with 50 μ l 50mM NH₄AcO; 20% MeCN (pH 6.2, adjusted with acetic acid), the fourth with 50 μ l 50mM NH₄AcO; 20% MeCN (pH 7.2), the fifth with 50 μ l 50mM NH₄HCO₃; 20% MeCN (pH 8.5) and the sixth with 50 μ l 0.1% NH₄OH; 20% MeCN (pH 9.5). 200 μ l of 0.5% acetic acid was added to each of the 6 fractions and they were subsequently desalted on C18 StageTips as previously described (Rappsilber et al., 2007) and evaporated to dryness in a vacuum concentrator. Peptides were reconstituted in 7 μ l 3% MeCN/0.1% formic acid (at an estimated concentration of 1 μ g/ μ l).

LC-MS/MS measurements

All peptide samples were separated on an online nanoflow EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific) and analyzed on a benchtop Orbitrap Q Exactive mass spectrometer (Thermo Fisher Scientific) as previously described (Mertins et al., 2013). Briefly, approximately 1 μ g of per peptide sample was injected onto a capillary column (Picofrit with 10 μ m tip opening / 75 μ m diameter, New Objective, PF360-75-10-N-5) packed in-house with 20cm C18 silica material (1.9 μ m ReproSil-Pur C18-AQ medium, Dr. Maisch GmbH, r119.aq). The UHPLC setup was connected with a custom-fit microadapting tee (360 μ m, IDEX Health & Science, UH-753), and capillary columns were heated to 50°C in column heater sleeves (Phoenix-ST) to reduce backpressure during UHPLC separation. Injected peptides were separated at a flow rate of 200 nL/min with a linear 80 min gradient from 100% solvent A (3% acetonitrile, 0.1% formic acid) to 30% solvent B (90% acetonitrile, 0.1% formic acid), followed by a linear 6 min gradient from 30% solvent B to 90% solvent B. Each sample was run for 150 min, including sample loading and column equilibration times. Data-dependent acquisition was

obtained using Xcalibur 2.2 software in positive ion mode at a spray voltage of 2.00 kV. MS1 Spectra were measured with a resolution of 70,000, an AGC target of $3e^6$ and a mass range from 300 to 1800 m/z . Up to 12 MS2 spectra per duty cycle were triggered at a resolution of 17,500, an AGC target of $5e^4$, an isolation window of 2.5 m/z and a normalized collision energy of 25. Peptides that triggered MS2 scans were dynamically excluded from further MS2 scans for 20 s.

MS identification and quantification of proteins

All mass spectra were analyzed with MaxQuant software version 1.3.5 (Cox and Mann, 2008) using the mouse UniProt database (March 2013). MS/MS searches for the proteome data sets were performed with the following parameters: Oxidation of methionine and protein N-terminal acetylation as variable modifications; carbamidomethylation as fixed modification. Trypsin/P was selected as the digestion enzyme, and a maximum of 3 labeled amino acids and 2 missed cleavages per peptide were allowed. The mass tolerance for precursor ions was set to 20 p.p.m. for the first search (used for nonlinear mass re-calibration) and 6 p.p.m. for the main search. Fragment ion mass tolerance was set to 20 p.p.m. The IBAQ feature was enabled in order to estimate relative proteins levels (Schwanhäusser et al., 2011). For identification we applied a maximum FDR of 1% separately on protein and peptide level. We required 2 or more unique/razor peptides for protein identification and a ratio count of 2 or more for protein quantification per replicate measurement.

Figures

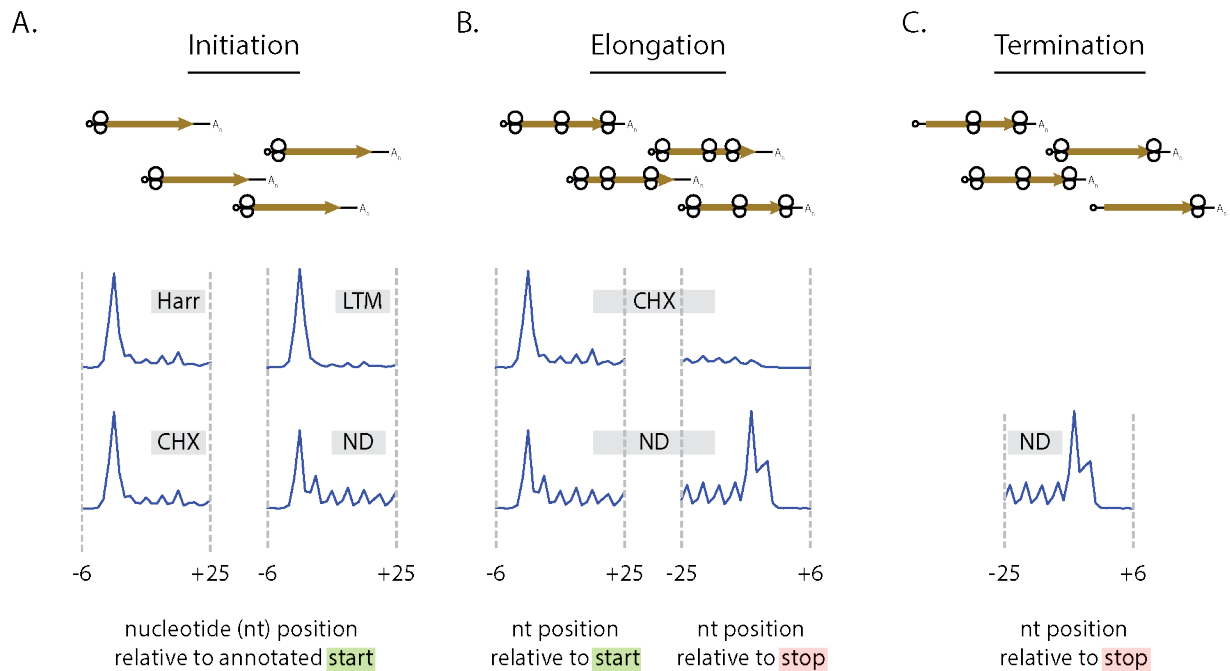


Figure 2-1. Features of translation observed in ribosome profiling data. (A) Ribosome profiling data from cells untreated (ND) or treated Harr, LTM, or CHX show peaks at translation initiation sites. (B) Ribosome profiling data from CHX or ND cells show phased elongation along the body of annotated ORFs. (C) Ribosome profiling data from ND cells show a peak at translation termination sites.

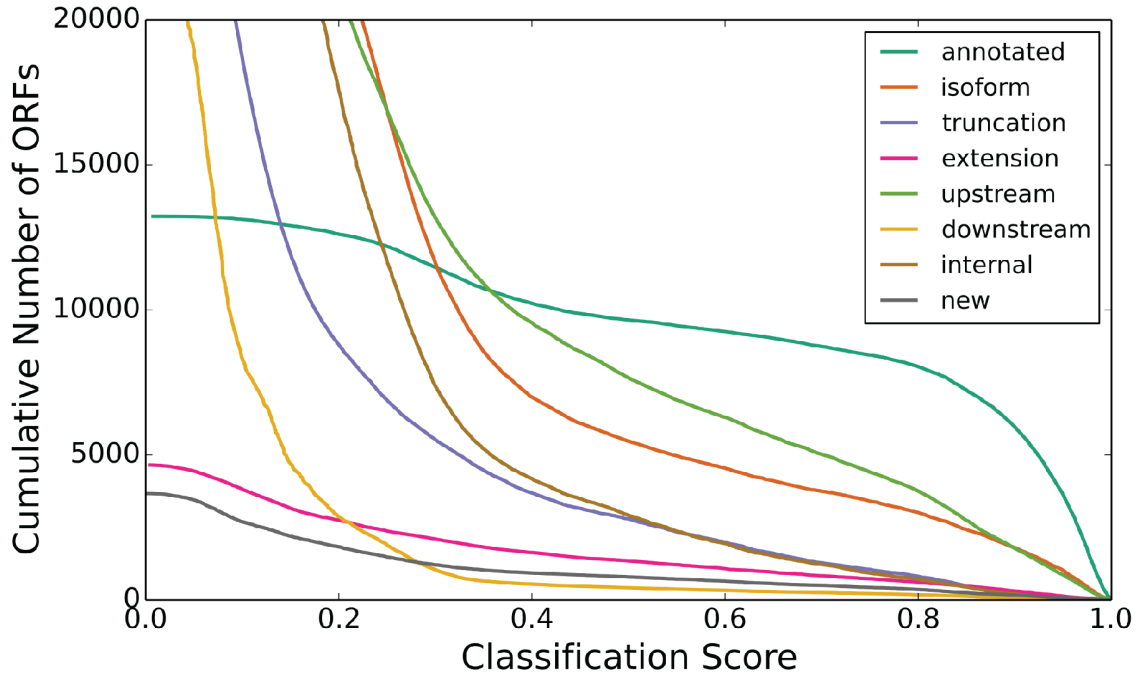


Figure 2-2. A final classification score captures known proteins and uncovers novel translation events. Outputs from our regression algorithm were combined via a Random Forest classifier to provide a single classification score. When a threshold is set at 0.8, our classifier captures most annotated proteins.

References

- Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses. *Science* *326*, 257–263.
- Chevrier, N., Mertins, P., Artyomov, M.N., Shalek, A.K., Iannacone, M., Ciaccio, M.F., Gat-Viks, I., Tonti, E., DeGrace, M.M., Clauser, K.R., et al. (2011). Systematic Discovery of TLR Signaling Components Delineates Viral-Sensing Circuits. *Cell* *147*, 853–867.
- Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* *26*, 1367–1372.
- Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals. *Molecular Cell* *47*, 810–822.
- Kim, D., Perte, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* *14*, R36.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Meth* *9*, 357–359.
- Lutz, M.B., Kukutsch, N., Ogilvie, A.L., Rössner, S., Koch, F., Romani, N., and Schuler, G. (1999). An advanced culture method for generating large quantities of highly pure dendritic cells from mouse bone marrow. *J. Immunol. Methods* *223*, 77–92.
- Mertins, P., Qiao, J.W., Patel, J., Udeshi, N.D., Clauser, K.R., Mani, D.R., 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 Meth* *10*, 634–637.
- Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of rRNA uncovers principles of rRNA production and degradation dynamics in mammalian cells. *Nat Biotechnol* *29*, 436–442.
- 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.
- Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* *473*, 337–342.
- Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R.,

Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498, 236–240.

Chapter 3

Results

Results

We used the regression-based approach described above to search our data for ORFs that display the three core features of translation that can be observed in ribosome profiling data: translation initiation, phased elongation, and translation termination. Our classification pipeline yielded 19,511 high confidence expressed ORFs (Fig. 3-1 A). 15,453 of these ORFs initiate with an ATG codon, and 4,058 initiate with other NTG codons. 41% (8,042) ORFs in our classified set map directly to previously annotated forms of proteins. 23% of our classifications correspond to variants of annotated proteins; we find 829 N-terminal truncations, 609 N-terminal extensions, and 3,129 isoforms. Our method predicts the expression of 6,902 novel ORFs. 472 of these are translated from RNAs that were previously not known to be protein-coding; 5,215 and 236 ORFs appear upstream and downstream, respectively, of annotated proteins; and 989 ORFs are internal and out of frame with respect to annotated ORFs. The majority of these novel ORFs are shorter than 100 codons (Fig. 3-1 B).

In order to confirm expression of our classifications, we performed LC-MS/MS measurements on LPS-stimulated BMDCs. Of the 8,042 annotated proteins that our algorithm classifies as being expressed, we measure peptides for 22.6%. We see uniquely-mapped peptides for 250 proteins in our newly classified set. The protein identification rate varies by ORF type; we identify peptides for 1% of the truncations, 1.7% of the extensions, 7.1% of the isoforms, 1.8% of the new, 0.11% of the uORFs, 4.2% of the dORFs, and 0.02% of the internals.

We classify 829 high confidence N-terminal truncation products; on average these truncations remove 21% of the annotated protein sequence. In absolute terms, these truncations result in proteins that are a median of 42 amino acids shorter (Fig. 3-2 A and B). 38.6% of these truncations initiate at ATG codons; 32.3% at CTG; 21.3% at GTG; and 7.7% at TTG (Fig. 3-2

C). 42 of these truncations lose their transmembrane domains (e.g., Fig. 3-2 D) and 89 of these truncations lose their ER-localizing signal sequences (e.g., Fig. 3-2 E), suggesting that these truncated proteins become localized to the cytosol rather than a membrane or the secretory pathway. 26 truncations lose their mitochondrial targeting domains.

We classify 609 high confidence N-terminal extension products; on average these extensions add 24.6% more length to the annotated protein sequences. In absolute terms, these extensions result in proteins that are a median of 22 amino acids longer (Fig. 3-3 A and B). 58% of these extensions initiate at CTG codons; 19.7% at GTG; 14% at TTG; and only 8.2% at ATG (Fig. 3-3 C). 25 of these extensions append a transmembrane domain to a previously cytosolic protein (e.g., Fig. 3-3 D) and 15 of these extensions gain an ER-localizing signal sequence (e.g., Fig. 3-3 E). 31 extensions add a mitochondrial targeting domain.

We classify 472 high confidence novel ORFs; the average and median lengths are 32 and 19 amino acids, respectively (Fig. 3-4 A and B). 72.7% of these novel ORFs initiate at ATG; 16.9% at CTG; 5.5% at GTG; and 4.9% at TTG (Fig. 3-4 C). 20 novel ORFs contain a transmembrane domain (e.g., Fig. 3-4 D and E), and 15 are predicted to encode an ER-localizing signal sequence (e.g., Fig 3-4 F). 3 novel ORFs are expected to contain a mitochondrial targeting sequence.

We classify 5,215 high confidence ORFs that initiate upstream of annotated ORFs; 4,249 (81.5%) of these are fully upstream, non-overlapping ORFs (hereafter referred to as ‘upstream’ or ‘uORFs’), and 966 initiate upstream of and terminate within annotated ORFs (hereafter referred to as ‘start-overlaps’). Upstream ORFs are of mean length 17.4 amino acids and median length of 11 amino acids (Fig. 3-5 A and B); and start-overlaps are of mean length 40.7 amino acids and median length of 32 amino acids (Fig. 3-6 A and B). 61.2% of uORFs initiate at ATG;

24.2% initiate at CTG; 8.5% at GTG; and 6.1% at TTG (Fig. 3-5 C). 43.2% of start-overlaps initiate at ATG; 35.1% initiate at CTG; 12.6% initiate at GTG; and 9.1% at TTG (Fig. 3-6 C). 30 uORFs contain predicted transmembrane domains (e.g., Fig. 3-5 D and E); 26 contain ER-localizing signal sequences (e.g., Fig. 3-5 F); and 57 contain a mitochondrial targeting sequence. 17 start-overlaps contain predicted transmembrane domains (e.g., Fig. 3-6 D and E); 26 contain ER-localizing signal sequences (e.g., Fig 3-6 F); and 58 contain a mitochondrial targeting sequence.

We classify 226 high confidence downstream ORFs; their average and median lengths are 21.1 and 11.5 amino acids, respectively (Fig. 3-7 A and B). 60.2% of downstream ORFs initiate at ATG; 29.6% initiate at CTG; 6.2% initiate at GTG; and 4.0% initiate at TTG (Fig. 3-7 C). 2 downstream ORFs contain predicted transmembrane domains (Fig. 3-7 D and E); 4 contain ER-localizing signal sequences (e.g., Fig 3-7 F); and 1 contains a mitochondrial targeting sequences.

We classify 989 high confidence internal, out-of-frame ORFs; the average and median lengths are 23.1 and 14 amino acids, respectively (Fig. 3-8 A and B). 78.8% of internals initiate at ATG; 10.9% initiate at CTG; 6.8% initiate at GTG; and 3.5% initiate at TTG (Fig. 3-8 C). 6 internal ORFs contain a predicted transmembrane domain (e.g., Fig. 3-8 D and E); 12 contain ER-localizing signal sequences (e.g., Fig 3-8 F); and 21 contain mitochondrial targeting domains.

For these classified ORFs, we quantified expression levels from ribosome profiling data and identify many whose expression changes significantly upon LPS stimulation. We recovered expression data from two CHX replicates and one ND sample; expression in these channels was compared to CHX data from a mock-stimulated experiment. ORFs displayed in Figure 3-9 show

LPS-induced changes across several orders of magnitude, but importantly, they show little induction or no expression in the mock samples. ORFs of all classes display this level expression change; those in Fig. 3-9 A-F show such changes in expression for new ORFs, Fig. 3-9 G for an upstream ORF, and Fig. 3-9 H for a start-overlap ORF. These newly classified ORFs are immediate candidates for knock-down/-out functional experiments.

Figures

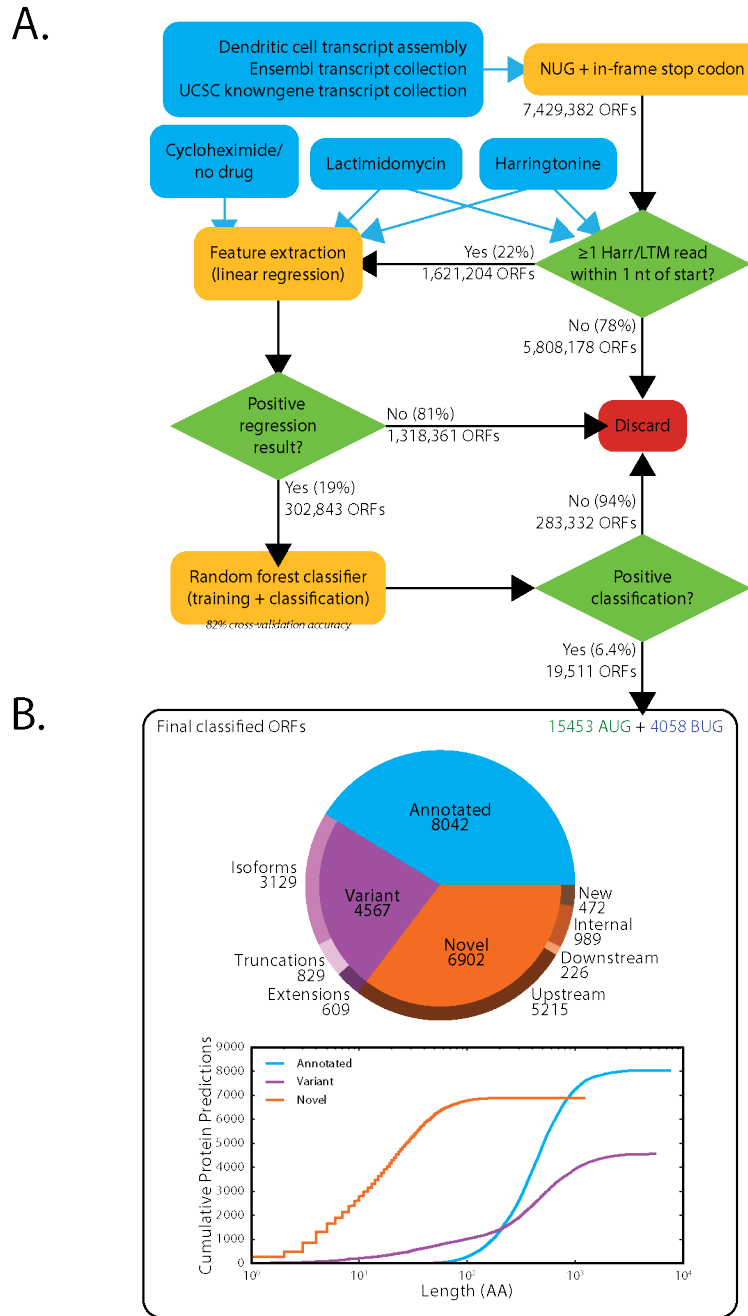


Figure 3-1. A CDS annotation pipeline and summary findings. (A) A pipeline that considers all possible ORFs on transcripts from the UCSC known gene set, Ensembl, and DC-specific transcript models. A regression-based approach (see text) is used to evaluate whether any of these ORFs are indeed being translated under various conditions (Harr, LTM, CHX, ND). Regression scores from all datasets for any one given ORF are combined via a random forest classifier. (B) This approach identifies many annotated and novel translated ORFs, the majority of which are shorter than 100 amino acids.

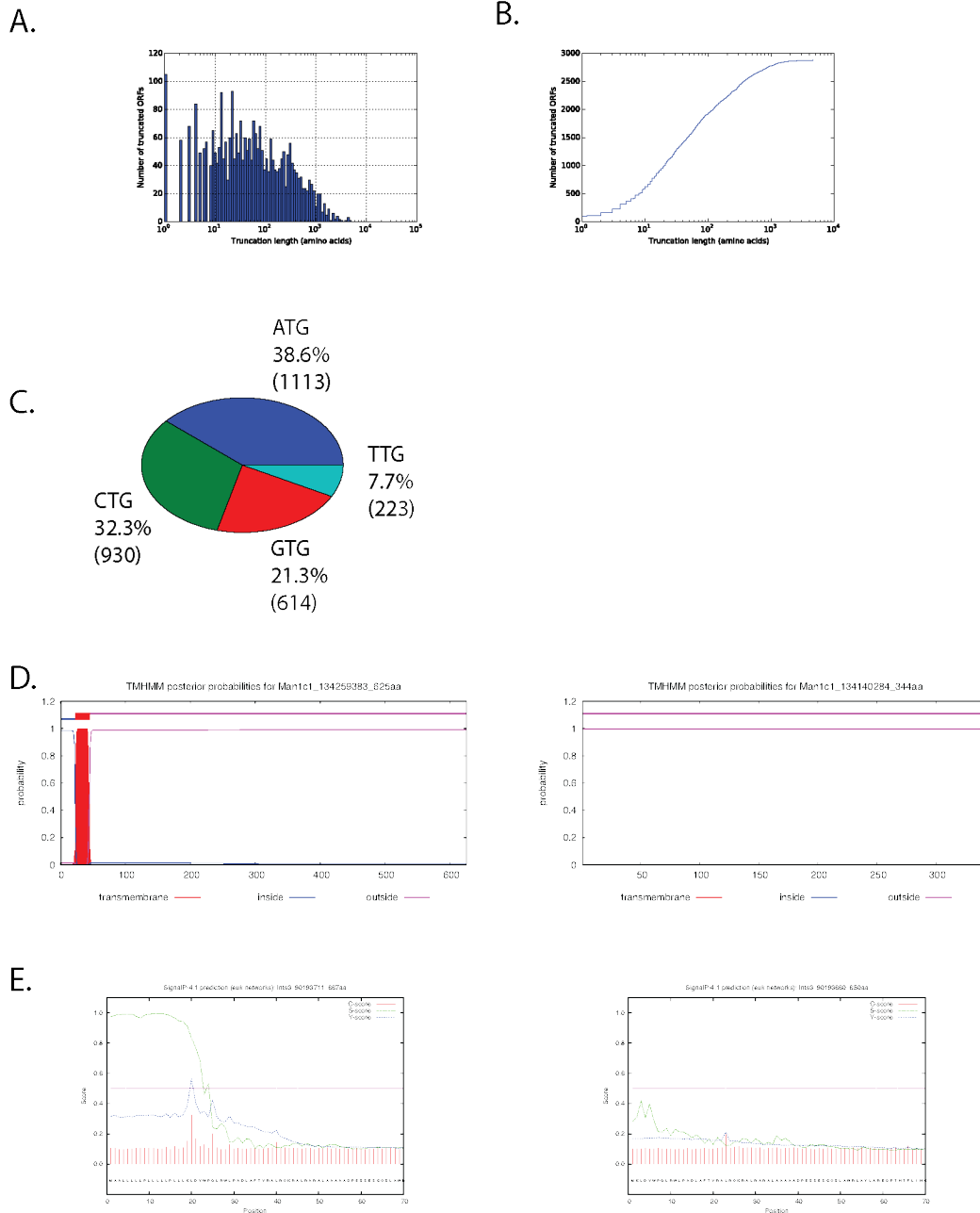


Figure 3-2. Truncated ORFs. (A) Size distribution in amino acids of truncated regions. (B) Cumulative distribution plot of data represented in (A). (C) Initiation codons for truncated products. Absolute numbers of ORFs from each category are given in parenthesis. (D) A truncation event that removes a transmembrane domain. On the left is the TMHMM plot for the full annotated sequence, and on the right appears the TMHMM plot for the truncated product that we classify. (E) A truncation event that removes an ER-localizing signal sequence. On the left is the SignalP plot for the full annotated sequence, and on the right appears the SignalP plot for the truncated product that we classify.

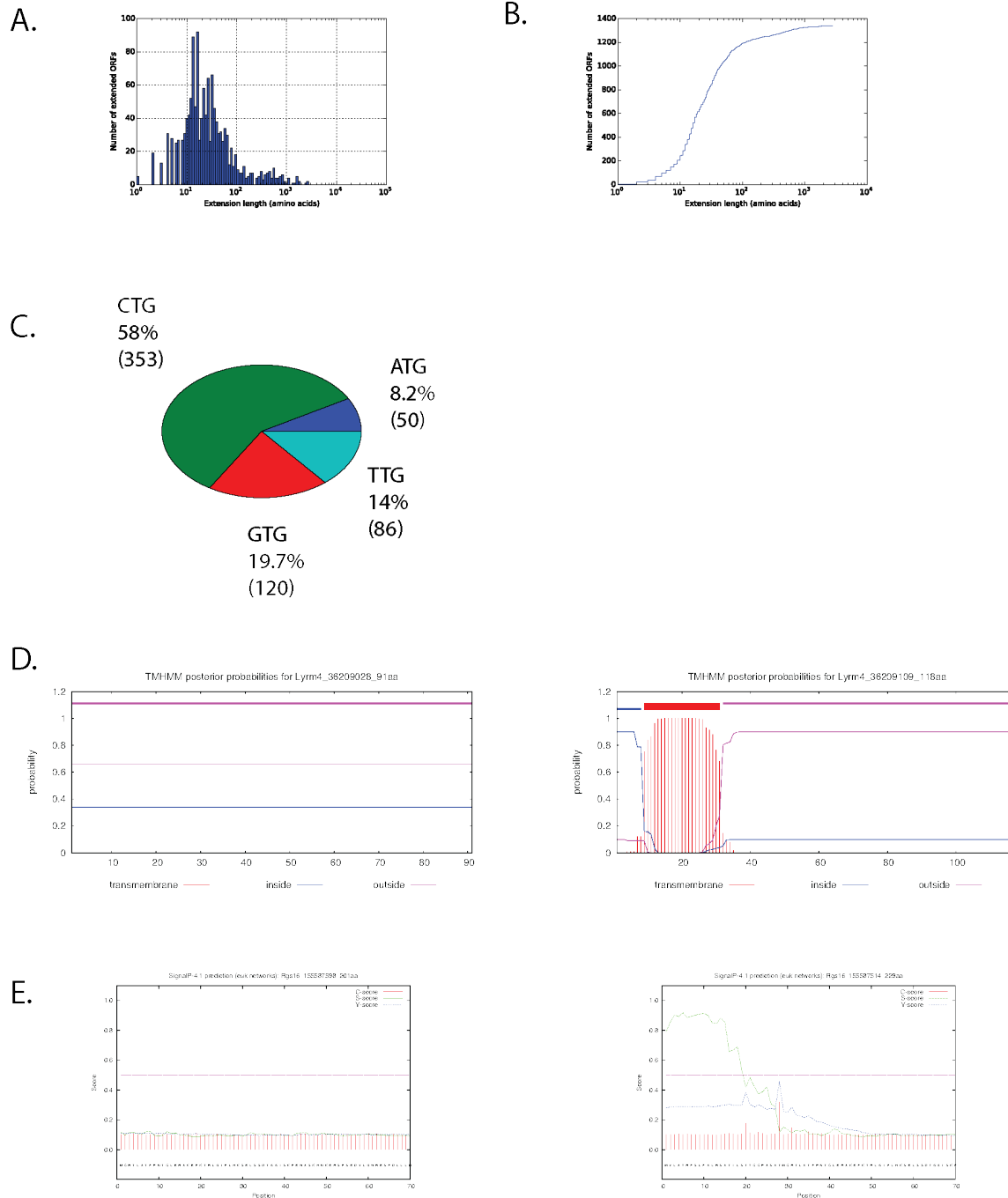


Figure 3-3. Extended ORFs. (A) Size distribution in amino acids of extended regions. (B) Cumulative distribution plot of data represented in (A). (C) Initiation codons for extended products. Absolute numbers of ORFs from each category are given in parenthesis. (D) An extension event that appends a transmembrane domain. On the left is the TMHMM plot for the full annotated sequence, and on the right appears the TMHMM plot for the extended product that we classify. (E) An extension event that appends an ER-localizing signal sequence. On the left is the SignalP plot for the full annotated sequence, and on the right appears the SignalP plot for the extended product that we classify.

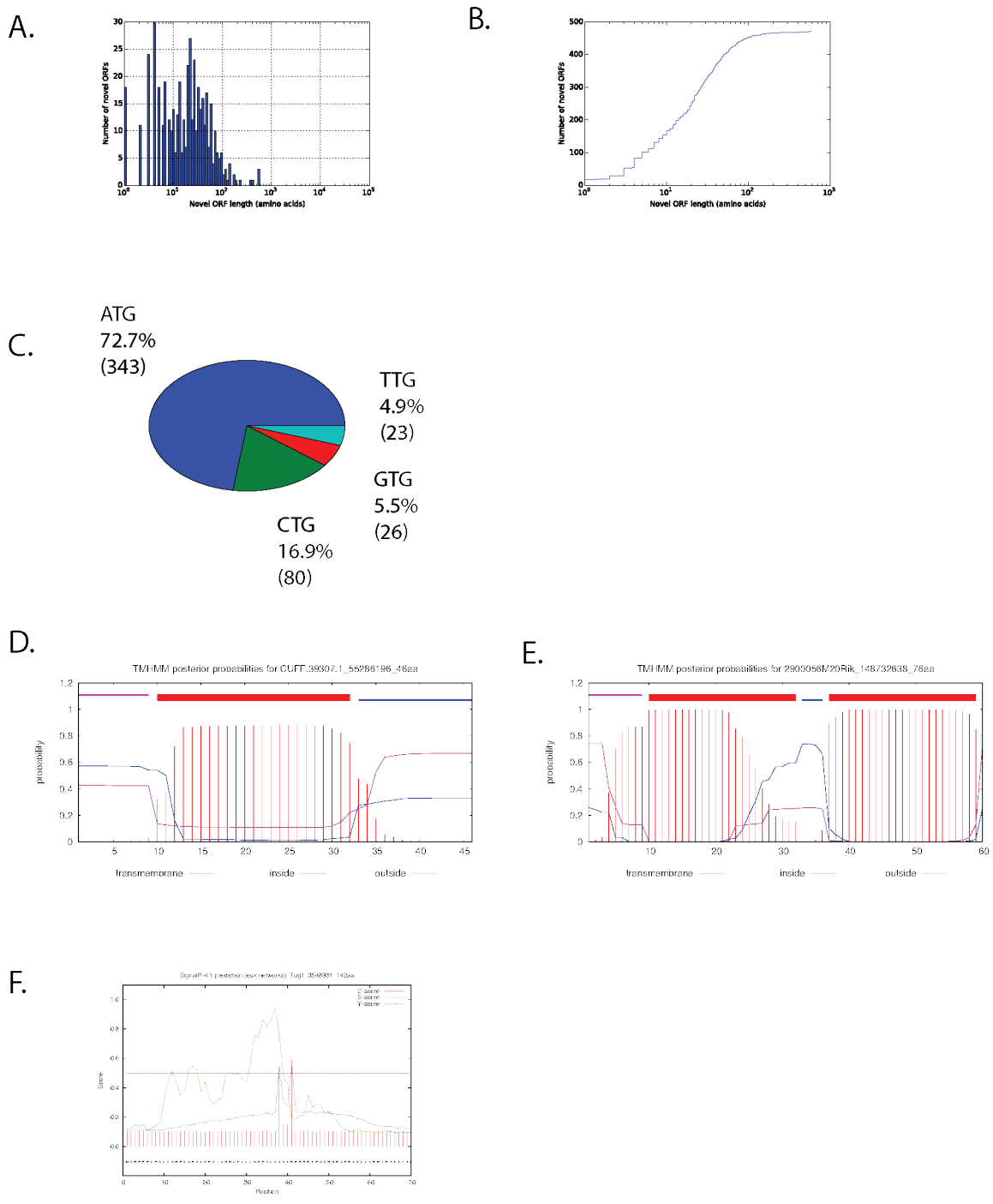


Figure 3-4. New ORFs. (A) Size distribution in amino acids of new ORFs. (B) Cumulative distribution plot of data represented in (A). (C) Initiation codons for new ORFs. Absolute numbers of ORFs from each category are given in parenthesis. (D) TMHMM plot for a new ORF with a predicted transmembrane domain. (E) TMHMM plot for a new ORF with two predicted transmembrane domains. (F) SignalP plot for a new ORF with a predicted ER-localizing signal sequence.

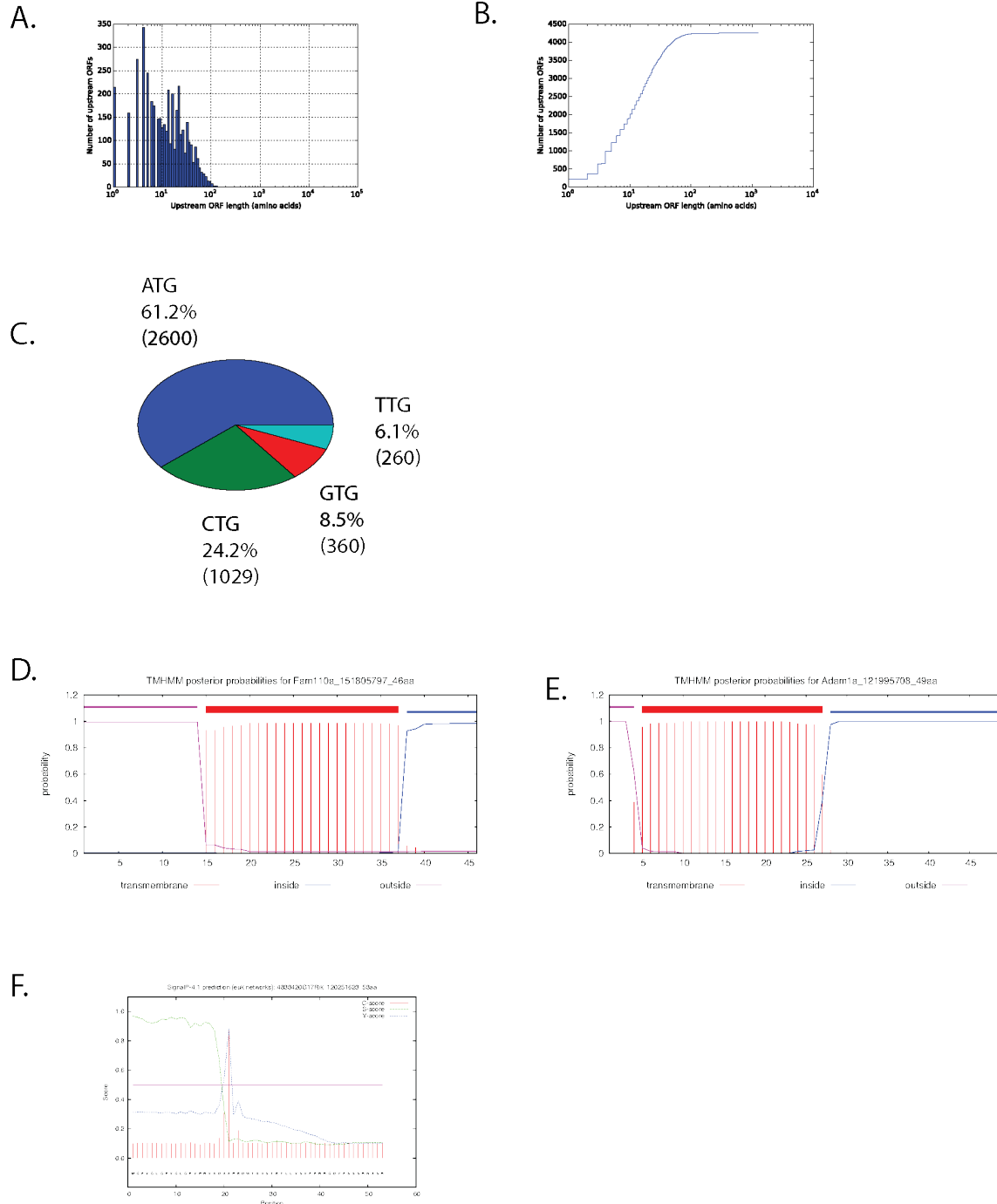


Figure 3-5. Upstream ORFs (uORFs). (A) Size distribution in amino acids of upstream ORFs. (B) Cumulative distribution plot of data represented in (A). (C) Initiation codons for uORFs. Absolute numbers of ORFs from each category are given in parenthesis. (D) TMHMM plot for a new uORF with a predicted transmembrane domain. (E) TMHMM plot for a new uORF with a predicted transmembrane domain. (F) SignalP plot for a new uORF with a predicted ER-localizing signal sequence.

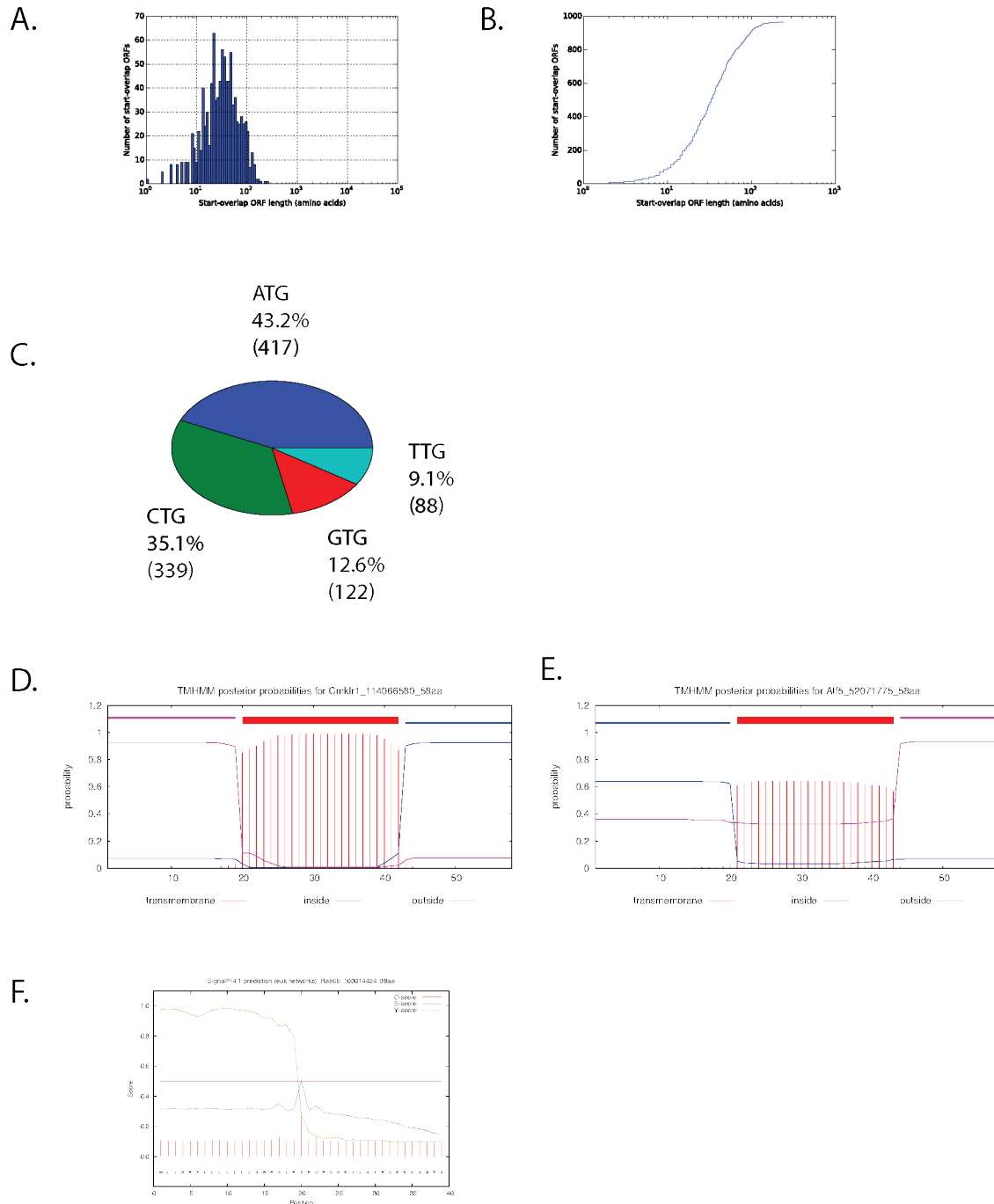
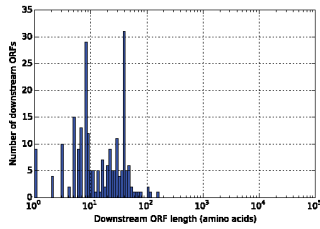
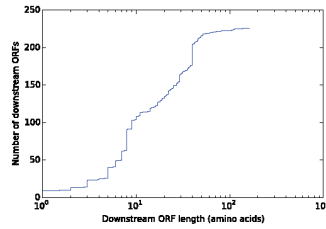


Figure 3-6. Start-overlap ORFs. (A) Size distribution in amino acids of start-overlap ORFs. (B) Cumulative distribution plot of data represented in (A). (C) Initiation codons for start-overlap ORFs. Absolute numbers of ORFs from each category are given in parenthesis. (D) TMHMM plot for a start-overlap ORF with a predicted transmembrane domain. (E) TMHMM plot for a start-overlap ORF with a predicted transmembrane domain. (F) SignalP plot for a start-overlap ORF with a predicted ER-localizing signal sequence.

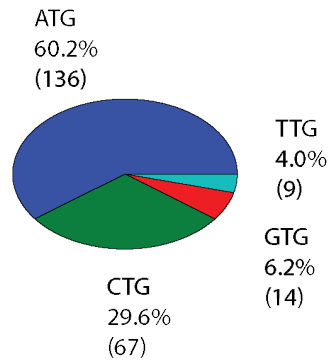
A.



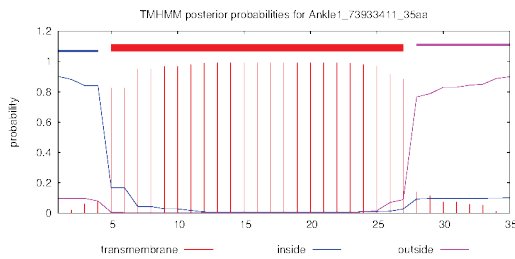
B.



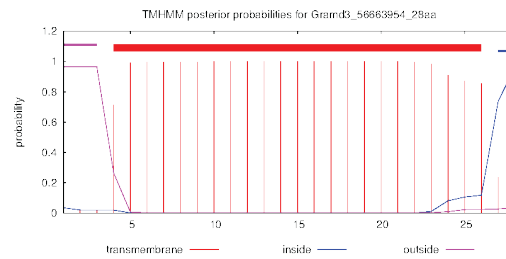
C.



D.



E.



F.

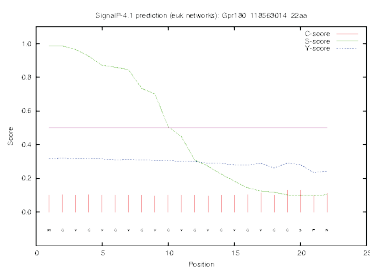


Figure 3-7. Downstream ORFs. (A) Size distribution in amino acids of downstream ORFs. (B) Cumulative distribution plot of data represented in (A). (C) Initiation codons for downstream ORFs. Absolute numbers of ORFs from each category are given in parenthesis. (D) TMHMM plot for a downstream ORF with a predicted transmembrane domain. (E) TMHMM plot for a downstream ORF with a predicted transmembrane domain. (F) SignalP plot for a downstream ORF with a predicted ER-localizing signal sequence.

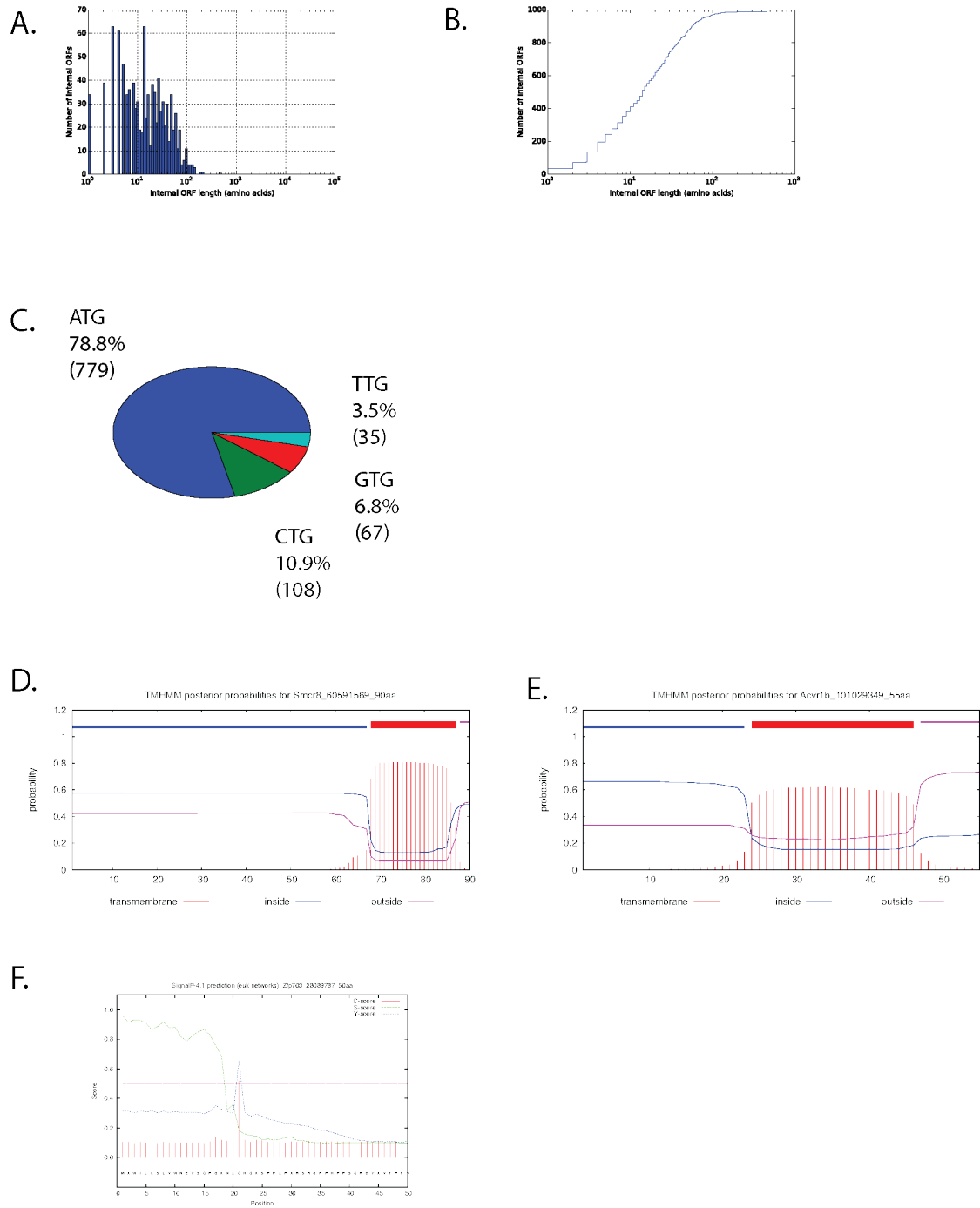


Figure 3-8. Internal ORFs. (A) Size distribution in amino acids of internal ORFs. (B) Cumulative distribution plot of data represented in (A). (C) Initiation codons for internal ORFs. Absolute numbers of ORFs from each category are given in parenthesis. (D) TMHMM plot for an internal ORF with a predicted transmembrane domain. (E) TMHMM plot for an internal ORF with a predicted transmembrane domain. (F) SignalP plot for an internal ORF with a predicted ER-localizing signal sequence.

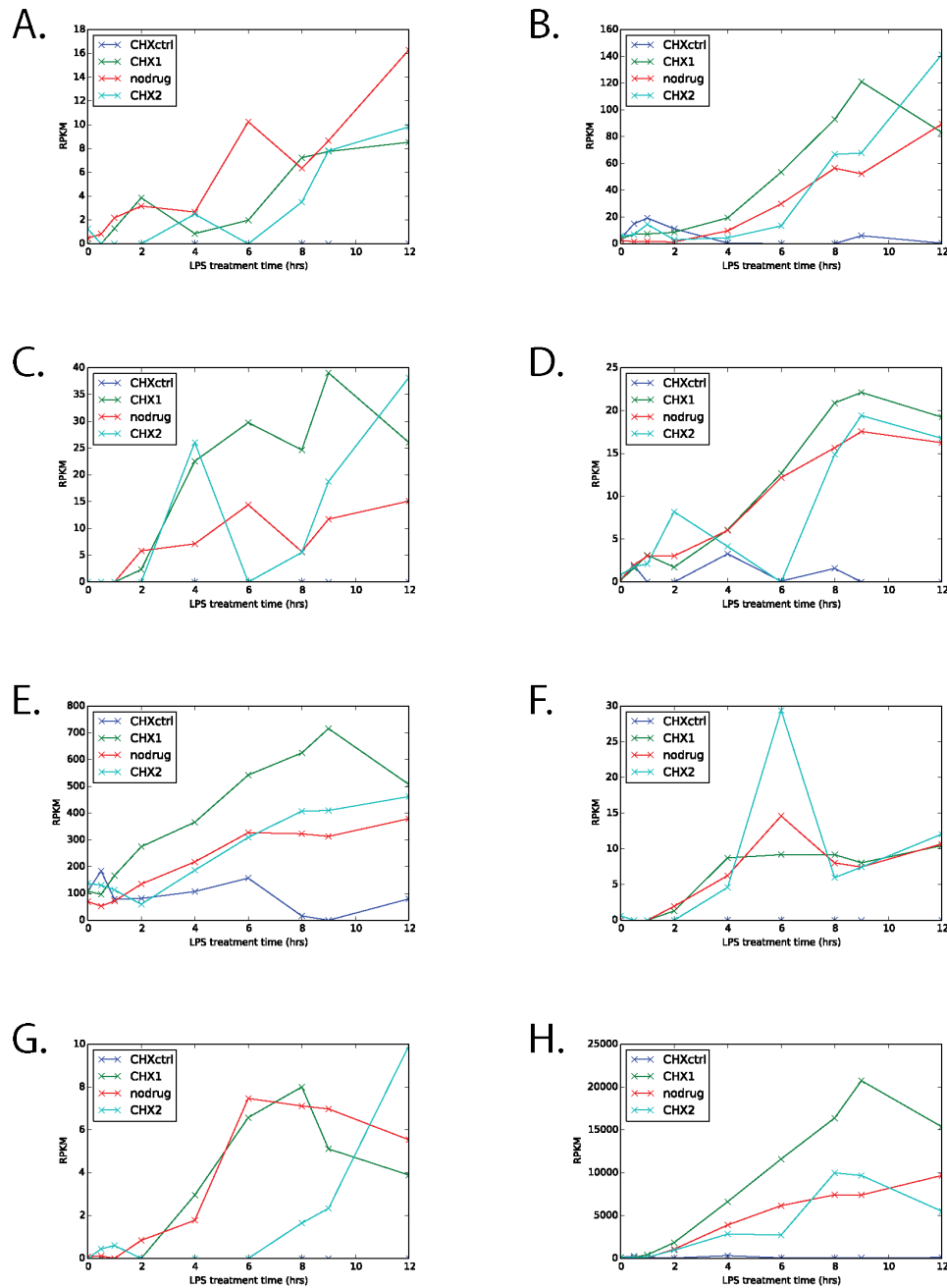


Figure 3-9. Expression changes in newly annotated ORFs upon LPS stimulation of DCs.

(A-F) Six new ORFs display a statistically significant increase in expression throughout the stimulation time-course relative to mock-stimulated cells. (A) Expression of CUFF.48463.1_103412190_37aa, (B) expression of Linc_lev2|NEW-CUFF.75079.1_11453242_236aa (C) expression of ANTI|Linc_lev2|NEW-CUFF.93902.1_87316357_121aa, (D) expression of Gm13822_115372103_10aa, (E) expression of Gm13822_115367507_38aa, and (F) expression of AK020236_84380027_45aa. (G) Tarm1_3503420_22aa is an upstream ORF that displays a statistically significant increase in expression upon LPS stimulation. (H) Ccl5_83343994_11aa is a start-overlap ORF that displays a statistically significant increase in expression upon LPS stimulation.

Chapter 4

Discussion and future perspectives

Discussion

Ribosome profiling provides a single-nucleotide resolution view of translation. From such high-resolution data, it is possible to extract the core features of translation: initiation, elongation, and termination. Here we describe an approach that uses all of these features of translation to annotate protein-coding sequences in a genome. We performed these experiments in LPS-stimulated BMDCs, a stereotyped context allows us to form hypotheses about the role of newly classified ORFs.

Since the original publication of the ribosome profiling approach, it has been possible to appreciate translation outside of canonical ORFs. Various methods have been introduced to separate dubious from bona fide translation. These methods have relied on single features of translation. This has been sufficient to identify highly-expressed, well-behaved genes. Our method takes into account all the possible features that can be gleaned from ribosome profiling in an effort to objectively and comprehensively catalog translated ORFs. In so doing, we identify ORFs that violate all the simple rules of ORF annotation pipelines past and present: we find translated ORFs that are short, that initiate at non-ATG codons, and that are completely overlapping and out-of-frame relative to annotated ORFs. We also identify N-terminal truncations and extensions of annotated ORFs, many of which lose or gain transmembrane or target sequences. Many of these newly annotated ORFs are indeed expressed and relatively long-lived, given that they are detectable by mass spectrometry.

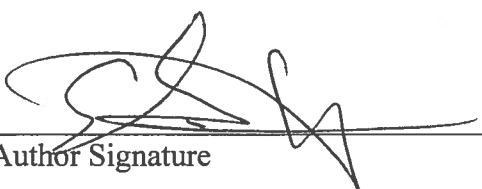
What we now know is that the universe of possible translation events is much larger than we had imagined, and that the cell breaks essentially all the “rules” of translation enshrined in the textbooks. Furthermore, these rules are broken in different ways for different cell types at different stages of development or activation. In the absence of simple heuristics, e.g., that

translation initiates at ATG codons only, the community of biologists needs a tool for the rapid and recurrent annotation of translation events. Our tool meets these criteria and will usher a more precise view of translation events and their resultant protein products.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

09/16/2014

Date