Developmental regulation of Canonical and small ORF translation from mRNAs

Pedro Patraquim[#1*,2], Muhammad Ali S. Mumtaz[#2], Jose I. Pueyo[2], Julie L. Aspden[3], and J.P. Couso[1*,2].

#: These authors contributed equally

1: Centro Andaluz de Biologia del Desarrollo, CSIC-UPO. Seville, Spain

2: Brighton and Sussex Medical School, Brighton East Sussex, UK

3: School of Molecular and Cell Biology, Faculty of Biological Sciences, University of Leeds LS2 9JT, UK

*: Current Address

Author for correspondence: J.P. Couso (jpcou@upo.es)

ABSTRACT

Ribosomal profiling has revealed the translation of thousands of sequences outside of annotated protein-coding genes, including small Open Reading Frames of less than 100 codons, and the translational regulation of many genes. Here we have improved Poly-Ribo-Seq and applied it to *Drosophila melanogaster* embryos to extend the catalogue of *in-vivo* translated small ORFs, and to reveal the translational regulation of both small and canonical ORFs from mRNAs across embryogenesis. We obtain highly correlated samples across five embryonic stages, with close to 500 million putative ribosomal footprints mapped to mRNAs, and compared them to existing Ribo-Seq and proteomic data. Our analysis reveals, for the first time in *Drosophila*, footprints mapping to codons in a phased pattern, the hallmark of productive translation, and we propose a simple binomial probability metric to ascertain translation probability. However, our results also reveal reproducible ribosomal binding apparently not resulting in productive translation. This non-productive ribosomal binding seems to be especially prevalent amongst upstream short ORFs located in the 5' mRNA Leaders, and amongst canonical ORFs during the activation of the zygotic translatome at the maternal to zygotic transition. We suggest that this non-productive ribosomal binding might be due to cis-regulatory ribosomal binding, and to defective ribosomal scanning of ORFs outside periods of productive translation. Finally, we show that the main function of upstream short ORFs is to buffer the translation of canonical ORFs, and that in general small ORFs in mRNAs display Poly-Ribo-Seq and bioinformatics markers compatible with an evolutionary transitory state towards full coding function.

KEYWORDS

Ribosomal profiling; Poly-Ribo-Seq; Ribosomal binding; Non-canonical translation; Regulation of translation; smORFs; sORFs; uORFs; maternal to zygotic transition.

BACKGROUND

Ribosomal profiling, the genome-wide sequencing of ribosome-protected RNA fragments (Ribo-Seq) has been increasing our understanding of a crucial event in all living genomes: protein translation [1, 2]. Ribo-Seq results show that there exists a noticeable level of non-canonical translation in eukaryotes. This can arise from non-

53 AUG codons, from polycistronic transcripts, and from unannotated small Open Reading
54 Frames of less than 100 codons (smORFs or sORFs) [3-5].

55

56 At the same time, the comparison of RNA-Seq and Ribo-Seq for the same biological
57 samples allows the transcriptome-wide study of an important but often forgotten
58 regulatory process in genome function: the regulation of translation. Typically,
59 expression of an mRNA transcript is equated with the automatic translation of any
60 encoded canonical ORF (of more than 100 codons). However, translation is not
61 automatic, but is regulated, often with great relevance for organismal function, both
62 under normal and altered conditions [6-10]. Translation regulation is also a key aspect
63 of development in all animals [11] and has been studied extensively in *Drosophila* [12-
64 14]. *Drosophila* embryogenesis is a highly-coordinated and complex process that is
65 completed in a time span of just 24 hours [15]. During the first two hours after egg
66 laying (AEL) there is absence of transcription from the zygotic genome and the key
67 developmental processes, such as establishment of the primary Antero-Posterior and
68 Dorso-Ventral axes, are controlled purely through the translational regulation of
69 maternal mRNA previously laid in the egg [16, 17]. After this initial period, the embryo
70 undergoes a ´maternal to zygotic transition´, whereby the transcription and translation
71 of the zygotic genome takes over the maternal products, a process also found in
72 nematodes, echinoderms, and vertebrates [13, 18-20]. Nonetheless, the impact of
73 translational regulation at the genome-wide scale on the whole of *Drosophila*
74 embryogenesis has not yet been revealed.

75

76 Ribo-Seq results regarding non-canonical and regulated translation have been the
77 subject of debate. While it has become accepted that both processes may occur more
78 extensively than previously thought, there is no consensus on the actual fraction of
79 smORFs and non-canonical ORFs whose translation is shown by Ribo-Seq [21-25].
80 The Ribo-Seq debate centres on the asymmetry between these numbers and other
81 translational evidence, and on the interpretation of the Ribo-Seq results themselves.
82 The most widely used counterpart of Ribo-Seq is proteomics, but the numbers of
83 protein and peptides detected by proteomics consistently fall short of those detected by
84 Ribo-Seq, especially regarding non-canonical translation. For example, the most
85 thorough proteomics study to date covering the whole *Drosophila melanogaster* life-
86 cycle has detected less than 40% of all unique canonical proteins [26]. This number is

87 further reduced to 30% of annotated smORF polypeptides, while we have previously
88 reported that 80% of canonical and small ORFs show clear Ribo-Seq evidence of
89 translation in a single embryonic cell line [23]. However, Ribo-Seq detects ribosomal
90 binding, not actual peptide production. There is not a universally agreed Ribo-Seq
91 metric unequivocally identifying productive, biologically-relevant translation, as
92 opposed to other processes such as low-level background translation, ribosomal
93 scanning and nonsense-mediated-decay surveillance, or stochastic ribosomal binding.
94 Bioinformatically, it is accepted that ribosomal binding above a certain level, and
95 especially, binding showing tri-nucleotide periodicity in phase with codon triplets
96 (phasing or framing), indicate translation of an ORF [1, 2]. A biochemical approach is to
97 introduce modifications to the ribosomal-RNA purification, to ensure that only
98 ribosomes engaged in productive translation are selected. For example Ribo-Seq of
99 polysomes (RNAs bound by several ribosomes), since the sequential translation of
100 polyadenylated, capped and circularised mRNAs by several ribosomes is a
101 supramolecular feature of productive translation, and excludes single ribosomes (which
102 could be involved in low-level translation but also in other activities) [23, 27]. We have
103 called this latter approach Poly-Ribo-Seq [23, 27].

104

105 Here we present an *in vivo* Poly-Ribo-Seq study covering a time-course of *Drosophila*
106 *melanogaster* embryogenesis. We have both improved our experimental Poly-Ribo-
107 Seq and the subsequent data analysis pipeline, to obtain unprecedented levels of Ribo-
108 Seq efficiency (reads mapped to ORFs) and quality, including codon framing as the
109 hallmark of productive, biologically meaningful translation. Thus, we can ascertain
110 translation and its regulation *in vivo* and across development for both canonical and
111 non-canonical ORFs. We detect the translation of thousands of non-annotated ORFs,
112 and identify hundreds of mRNAs whose translation is highly regulated during
113 embryogenesis. However, our results also reveal reproducible ribosomal binding not
114 resulting in productive translation. This non-productive ribosomal binding seems to be
115 especially prevalent amongst upstream short ORFs located in the 5' mRNA Leaders,
116 and amongst canonical ORFs during the activation of the zygotic translatome at the
117 maternal to zygotic transition. We suggest that this type of ribosomal binding might be
118 due to either cis-regulatory ribosomal activity, or to defective ribosomal scanning of
119 ORFs outside periods of productive translation.

120

121 **RESULTS**

122 **1) The method and overall data**

123 Since Poly-Ribo-Seq requires even larger amounts of starting material than Ribo-Seq
124 due to polysome fractionation, the *Drosophila S2* cell line was an excellent tool for the
125 development of the technique. However, the S2 cell line is derived from just one type of
126 tissue (macrophage-like) from late stage *Drosophila melanogaster* embryos [28] and
127 may not be the ideal system to obtain a comprehensive picture of *Drosophila*
128 translation. Only 60% of both canonical genes and uORFs, and only about a third of
129 annotated smORFs (hereafter referred as short CDSs, or sCDSs [29]) appear
130 transcribed in this cell line [23]. Therefore, we applied Poly-Ribo-Seq to *D.*
131 *melanogaster* embryos in order to extend the catalogue of *in-vivo* translated smORFs
132 and study translation across three distinct developmental time windows of *Drosophila*
133 embryogenesis. The original protocol from Aspden [23] was adapted and improved in
134 terms of biochemical rRNA depletion and bioinformatics mapping of FPs to ORFs (see
135 methods), and we obtained higher yield, purity and accuracy (Fig.1 and Sup. Table 1).
136 This way we detected rare but reproducible translation events in non-annotated genes,
137 and obtained a comprehensive picture of translation across development.

138

139 In our experimental design, we divided the *Drosophila* embryogenesis in three temporal
140 8 hour windows: Early (0-8 hours after egg laying, or 0-8h. AEL), Mid (8-16h. AEL) and
141 Late embryogenesis (16-24h. AEL) (Fig. 1A). The key developmental processes
142 occurring during these periods of embryogenesis are described elsewhere [15]. Briefly,
143 early embryogenesis (0-8h. AEL) is characterised by the maternal to zygotic transition,
144 followed by gastrulation, germ band formation and ectodermal and mesodermal
145 determination and pattern formation, including the determination of the CNS. The 8-
146 16h. AEL hours of development (mid-embryogenesis) covers the process of tissue
147 formation and organogenesis (such as presumptive gonad and gut assembly) including
148 multi-organ processes such as formation of segments and the feeding cavity. The
149 differentiation of neurons and muscles also occur during this time. Late embryogenesis
150 (16-24h. AEL) is characterised by final differentiation and the start of physiological
151 processes such as epidermal and cuticle differentiation [30] (with opening of tracheae
152 to respiration) and digestive system differentiation (with yolk digestion and uric acid)

153 [31, 32], plus the connection and fine-tuning of neural, sensory and muscle structures,
154 eventually resulting in coordinated embryo movements and larval hatching [33, 34].

155

156 We collected two biological replicas (called T and B) from each embryonic stage,
157 extracting simultaneously total RNA and Ribosomal Footprints. Sequencing and
158 analysis of these samples was used to determine transcription and translation levels,
159 which in turn were used to reveal the extent and quality of translational regulation
160 during *Drosophila* embryogenesis at a genomic scale. We studied several classes of
161 ORFs as defined in Couso and Patraquim [29], altogether numbering 40,852 ORFs
162 (see methods): 21,118 canonical ORFs of more than 100 codons in polyadenylated
163 mRNAs; 862 ORFs from annotated coding sequences of 100 codons or less in an
164 otherwise canonical mRNA (short CDSs); and 18,872 upstream ORFs (uORFs) in the
165 5' leaders of canonical mRNAs. We studied only AUG-START ORFs, and also
166 discarded overlapping uORFs, in order to ensure the accuracy of our translation
167 assessments. Similarly, the translation of ORFs in putative long-non-coding RNAs
168 (lncORFs) will be described in a forthcoming study due to its special characteristics.

169

170 As the Ribo-Seq protocol typically results in a composite library with a majority of
171 ribosomal rRNA reads, and a minority of ribosomal footprints (FPs), we sequenced an
172 exhaustive amount, totalling 1.2 billion Ribo-Seq reads, yielding some 345 million
173 ribosomal footprints (FPs) mapped to the genome (Sup. Table 1) with ~80% FPs
174 aligned to annotated CDS, ~16% to 5' Leaders and 6% to 3'UTRs (Figs.1B and 2A),
175 providing a 300x coverage or our 40,852-strong ORF-ome. The data shows high
176 reproducibility, with high Spearman correlation(r=0.9) of RPKMs (number of sequenced
177 reads per ORF length in kilobase, per million reads) for each ORF amongst the two
178 replicas (Fig. 1C) for both RNA-Seq ($RPKM^{RNA}$) and Ribo-Seq ($RPKM^{FP}$) samples. Only
179 reads of 26 to 36 nucleotides (nt) were counted for $RPKM^{FP}$, as this length distribution
180 corresponded to 97% of the reads obtained in all cases, and is within the size range
181 previously described for Ribosomal FPs [1] (Sup. Fig 1A). At first glance, our data
182 reveals translation of non-canonical genes, as for the polycistronic smORF gene *tarsal-*
183 *less* [35] (Fig. 1D and Sup. Table 3A).

184

185 2) **Framing: an indicator of productive translation**

186 For our bioinformatic analysis (see Methods and Fig. 2A), we use $RPKM^{RNA}$>1 to
187 indicate transcription of an ORF (in either of our two replicas). To determine translation,
188 we use two combined filters. First, a quantitative filter: a minimal amount of ribosomal
189 binding must be detected, set as a standard $RPKM^{FP}$>1 in both T and B replicas, to
190 indicate signal above error and background. Second, a qualitative filter, the nature of
191 that binding: the ribosomal movement across an ORF during translation generates tri-
192 nucleotide phasing of ribosomal protected footprints, or framing (Fig. 2B), in contrast
193 with the reads derived from RNA-Seq, which accumulate evenly across all nucleotides
194 in mRNA (Sup. Fig 1B).

195

196 Previous ribosomal profiling studies of *Drosophila* had not been able to observe strong
197 framing on a genome-wide scale, and thus translation has been proposed on the basis
198 of RPKM and other quantitative metrics, such as coverage [18, 23, 24, 36]. However,
199 our improved Poly-Ribo-Seq method, coupled with an improved bioinformatics analysis
200 of the data (see methods), revealed framing for the first time in the *Drosophila* (Figure
201 2B-C). Like all Ribo-Seq experiments, our data revealed a distribution of Ribosomal
202 footprint lengths (Sup. Fig 1A). We focused on the most represented read length (32
203 nucleotides), which also shows the clearest framing in all samples (55-60% reads in-
204 frame; Figure 2C). Different formulae and metrics have been previously used to
205 quantify framing per ORF and provide a cut-off [1, 21], reviewed in [5, 37], which was
206 usually obtained by sampling a small number of bona-fide translated ORFs and
207 extrapolating the results. While these methods can identify translation, they can be
208 complex and difficult to apply universally to all Ribo-Seq datasets, due to differences in
209 sampling and protocols utilised.

210

211 Here, we have opted for a simpler approach that is easy to extend to past and future
212 datasets. We treat framing as a simple coin-toss problem, which offers clear statistical
213 probabilities, even with a low number of reads. As each footprint is the result of an
214 independent RNA-ribosome molecular interaction, each resulting sequencing read has
215 in principle a 1/3 probability of aligning to the right frame in an ORF, and 2/3 of aligning
216 to the wrong frame (Fig. 2D). The probability of having *x* successful outcomes (reads
217 aligned to the ORF frame) following *n* independent trials (number of reads mapping to
218 the ORF) follows a classical binomial distribution, which can be consulted using widely
219 available tables [38] and computer programs. We take the observed binomial

220 probability as the probability of being translated for a given ORF and we consider it as
221 indicative of translation when p<0.01 (i.e. the probability of observing such framing of
222 reads in the ORF by chance is less than one in a hundred; see Methods for details and
223 implementation). Here, because of our $RPKM^{FP}$>1 filter, and because only 32 nt. FP
224 reads were used to calculate framing (see methods and Sup.Fig.1) we noted that, out
225 of 44,067 translation calls in all stages studied, only 36 calls appeared translated
226 having less than 20 ribosomal footprints (of 26 to 36nts) and only 2 calls had less than
227 10 footprints. As example (Fig. 2E and Sup. Table 3; see also Sup. Data for full sets of
228 data) we show the non-canonical smORF gene *sarcolamban* (*scl*) whose translation
229 has been determined experimentally [39]. Its two ORFs (*SclA* and *SclB*) achieve
230 $RPKM^{FP}$ above 1 in both T and B replicas during Mid and Late embryogenesis. Most of
231 its FP reads mapped to frame "2", but *SclB* only achieved a binomial *p*-value below
232 0.01 during Late embryogenesis. Thus, our data indicated that *SclA* is ribosomal-bound
233 and translated at Mid- and Late embryogenesis, but *SclB* is only productively translated
234 during Late embryogenesis (when *scl* function in regulating muscle contraction [39]
235 would have its maximum requirement, see above and Fig.1A). Note that framing *p*-
236 values do not indicate the amount of translation (which is indicated by the $RPKM^{FP}$ and
237 its associated metric translational efficiency, TE, see below) but only the likelihood that
238 the observed ribosomal binding leads to productive ORF translation. This binomial
239 framing metric also ensures that the translation detected belongs to a given ORF, not
240 to another putative ORF overlapping it.

241

242 Using these criteria, we observed that 98% of transcribed canonical ORFs, 90% of
243 short CDSs and 71% of uORFs were reproducibly bound by ribosomes at any point
244 during development (showed $RPKM^{FP}$>1 in both replicates of the same stage),
245 whereas the percentage of actually translated ORFs (ribosomal-bound ORFs also
246 having framing p-value<0.01 in that stage) was 92% of Canonical, 72% of short CDSs,
247 and 12% of transcribed uORFs (Fig. 2F).

248

249 **3) Consistency of Transcription and Translation measurements across different**
250 **types of genomic data**

251 We further verified the robustness and reproducibility of our data and analysis by cross-
252 comparisons across related gene-expression genomic data: RNA-Seq, Ribosomal

253 profiling, and quantitative proteomics. For RNA expression, we consolidated the
254 modENCODE embryo RNA-Seq data [40], into our three 8-hour long developmental
255 time windows. For all three developmental windows, the $RPKM^{RNA}$ of RNA-Seq
256 samples was more highly correlated (in the region of *r*=0.9) with a replica of their stage
257 (either ours or modENCODE), than with samples of other stages (Fig. 1C, Fig. 3A, Sup
258 Table 2).

259

260 For Ribosomal profiling, our $RPKM^{FP}$ data showed an even more marked trend to
261 correlate preferentially first with their own stage replica (r=0.98-0.99), and second with
262 samples from adjacent time periods (Fig. 1C, 3A, Sup Table 2). Our FPs and RNA-Seq
263 from the same stage showed a high correlation of r=0.8-0.9 (Fig. 1C, Sup Table 2). We
264 also compared our FP data with the Kronja et al. 2014 Ribo-Seq data from unfertilized
265 *Drosophila* eggs, that was obtained after RNaseI digestion akin to our own protocol. As
266 expected, the $RPKM^{FP}$ from unfertilized eggs (FP Maternal) showed the highest
267 correlation with our $RPKM^{FP}$ samples from early embryogenesis (0-8h AEL)
268 (Spearman´s *r*≈0.8) (Fig. 3A and Sup. Table 2). Our bioinformatic pipeline detected
269 framing in Kronja´s ORF-mapping reads of 29nt (see methods). 10,314 canonical
270 ORFs appeared translated both in unfertilized eggs and Early embryos, albeit showing
271 increased translation efficiency in the latter (Fig. 3B). However, differences between
272 the maternal and zygotic translatomes were also identified, since the Kronja [18] data
273 only assessed maternal RNA translation while our 0-8h AEL sample detects both
274 maternal and early zygotic translation (see “Regulation of translation” below).

275

276 We also compared our embryo FP data with *Drosophila* embryo-derived S2 cell
277 cultures [28], which in principle should provide synchronic data less subjected to
278 experimental and developmental noise. We have re-analysed the data from Aspden et
279 al. 2014, adding a new Poly-Ribo-Seq sample obtained here. S2 translation in general
280 is highly correlated with Early embryos (Fig. 3A), and the proportions of canonical
281 ORFs ribosome-bound and translated are 98.0% and 72%, respectively. Despite this
282 overall similarity, differences were also noted (see “Regulation of translation” below).

283

284 For proteomics, we consolidated the results of extensive quantitative proteomics
285 across embryogenesis [26] in our three 8h developmental time windows (methods).

286 This quantitative proteomics data showed an overall 0.45 correlation with our $RPKM^{FP}$
287 values (Fig. 3C). This correlation appears modest, but it is highly significant (at
288 $p<0.0001$) indicating that $RPKM^{FP}$ can be a good proxy of protein-producing, productive
289 canonical translation. However, 40% canonical ORFs detected as transcribed by RNA-
290 Seq by us and modENCODE during embryogenesis are not detected by proteomics
291 (Fig. 3D), while this number is reduced to 8% by our Poly-Ribo-Seq pipeline. The
292 median $RPKM^{FP}$ value of the proteomics-detected proteins (32.11) is much higher than
293 for those not detected (13.41), indicating that proteomics detection needs high levels of
294 protein translation, as noticed previously [21, 23]. Yet, many ORFs with experimentally-
295 verified translation, and with high Poly-Ribo-Seq metrics, are not detected by
296 proteomics, including all *Hox* genes except the more widely-expressed *Ubx* and *abd-A*.
297 This `proteomics detectability´ issue is not attributable to trypsination treatments, since
298 we observed that 99% of canonical and 92% of non-canonical ORF could produce K-
299 and R-cleaved peptides of 7 to 24aa-long suitable for mass-spec (MS) detection.
300 Hence, other yet unknown factors must influence the proteomics detection of proteins.
301 In addition, small peptides of less than 50 amino acids suffer from enhanced
302 degradation [41, 42], further hindering their detection by proteomics. Indeed, we
303 observed that sCDSs are especially handicapped for MS detection, since the median
304 $RPKM^{FP}$ of those detected is 172.1 (Fig. 3E). Finally, when comparing MS data with
305 RNA-Seq and Poly-Ribo-Seq, and contrary to that observed between RNA and FP data
306 (Fig. 3A) the highest correlation observed is amongst MS data from different embryonic
307 stages. Altogether our analysis suggests that proteomics is an appropriate technique to
308 detect constitutively and highly translated proteins, but not so useful for the detection of
309 ORFs translated either moderately, or in a developmentally-regulated manner.

310

311 smORFs also appeared harder to detect by Poly-Ribo-Seq than canonical ORFs,
312 requiring a median $RPKM^{FP}$ of 40.5 (Fig. 2F, 3E). As an independent indicator of
313 translation, we have tagged a sample of smORFs from sCDSs and added to the results
314 from Aspden [23] and [43], for a total of 44 sCDSs. We compared the results of tagging
315 with our Poly-Ribo-Seq pipeline, and with S2 cell proteomics data [44]. Out of the 44
316 tagged sCDSs, we observed that 43 were detected as translated by tagging, 29 by
317 Poly-Ribo-Seq, and 14 by proteomics (Sup. Fig. 2A). These results showed a coherent
318 pattern, where the three techniques, based on very different detection technologies
319 (Confocal microscopy vs. NextGenSeq vs. Mass-spec, respectively), show decreasing
320 sensitivity thresholds: Tagging > Poly-Ribo-Seq > Proteomics, as also noted when

321 comparing results from other species [45]. Thus, 12 of the 14 proteomics-detected
322 sCDSs were also detected by Poly-Ribo-Seq, and 28 of the 29 detected by Poly-Ribo-
323 Seq were also detected by Tagging. However, our expression of tagged peptides in S2
324 cells is based on non-endogenous high transcription and this might have pushed the
325 peptide production of lowly expressed ORFs over the edge of detection (see methods
326 and Aspden [23]). Hence, our tagging results may indicate what "can be translated"
327 rather than what actually "is". Accordingly, 11 of the 13 tag-positive sCDSs with no
328 Poly-Ribo-Seq evidence in S2 cells were translated in other stages. The most
329 abundant tagged peptide localisations (Fig. 3F) were a) reticular-punctate in the
330 cytoplasm, which may correspond to mitochondria [23], or else to mitochondria-
331 associated ER producing the peptide; and b) cortical, which could indicate and
332 association with the plasma membrane or its cytoskeletal cortex. These observations fit
333 with the postulated tendency of sCDSs peptides for membrane-associated localisations
334 [23, 29].

335

336 **4) Regulation of canonical and sCDS translation during embryogenesis**

337 According to our data and modENCODE, 82% of canonical ORFs and 57% of short
338 CDSs were transcribed at any point during development (Fig. 2F). Of these, 8% and
339 28% respectively were never translated during embryogenesis, giving a first indication
340 of the extent of translational regulation in *Drosophila melanogaster*.

341

342 Our data also showed that translation of some 68% of canonical ORFs was detected
343 constitutively (at all embryonic stages), whereas 32% were detected only at specific
344 stages of development (Fig. 4A). Stage-specificity seemed to be higher for smORFs,
345 since only 56% of short CDSs and 18% of uORFs were detected across
346 embryogenesis, whereas 44% and 82% respectively were translated only during some
347 stages of embryogenesis. However, this developmental analysis reflected both
348 transcriptional and translational regulation. To further quantify the extent and dynamics
349 of purely translational regulation, we analysed the changes in translational efficiency
350 (TE) [2] across stages (TE indicates the ratio between FPs and RNA, calculated as
351 $RPKM^{FP}$ / $RPKM^{RNA}$; see methods). This TE analysis revealed both on-off translation
352 changes, and modulations of sustained translation, from one embryonic stage to the
353 next. The average TE varied somehow across development (Fig. 4B), but in general

354 we observed fairly stable and correlated TE values across development (but see also
355 “uORFs as translational regulators”). To identify which ORFs undergo statistically
356 significant translational changes we used the Z-ratio method (which basically indicates
357 whether the variation of a given ORF TE is significantly outside the standard deviation
358 of the total sample; see methods) [46-48]. We also included in the analysis the
359 Unfertilized Egg and S2 cells TE data (comparing them to Early and Late
360 embryogenesis, respectively). We observed significant translational changes in all ORF
361 classes, with the percentage of changed ORFs depending on the ORF class. About
362 12% of canonical proteins (1,899 of 15,956 analysed), 16% of short CDSs (57 of 355)
363 and 18% of translated uORFs had significantly changed TE from any given stage to the
364 next, indicating significant translational regulation (Fig. 4C). One of the most abundant
365 changes was the translational up-regulation of 539 canonical ORFs from Early to Mid-
366 embryogenesis (Fig. 4D). These ORFs were enriched for GO expression terms
367 indicating a role in organogenesis (i.e. CNS, epidermis and digestive system
368 development) (Fig.4E), as befits the developmental processes underway (Fig. 1A).

369

370 The comparison of unfertilized eggs with 0-8h AEL embryo data revealed the maternal
371 to zygotic transition in the translatome. The degradation of maternal mRNAs after the
372 zygotic to maternal transition from 2h AEL, and their substitution by zygotically-
373 transcribed mRNAs has been reported, but a number of mRNAs do persist during early
374 embryogenesis [49]. Our results suggest that some of these maternal mRNAs that
375 persist are subjected to translational repression as an added layer of gene regulation,
376 whereas others go on to increase their translation during embryogenesis. Applying Z-
377 ratio analysis, we observed in total 374 canonical ORFs with significantly changed
378 translational efficiency between the maternal and the early zygotic translatomes (Fig.
379 4C), of which 237 where down-regulated (including 58 *Dscam* ORFs) and 137 up-
380 regulated, showing RNA expression patterns preferentially registered as ´maternal´ or
381 ‘ubiquituous´ at this stage (Fig. 4D,F). Further, the maternal to zygotic transition in
382 translation was also highlighted by the correlation between qualitative and quantitative
383 changes in ribosomal binding. 259 maternal-only canonical ORFs appeared translated
384 (framed) in Eggs but not in 0-8h AEL embryos (Fig. 3B), undergoing a significant
385 reduction in ribosomal binding efficiency during this period (as shown by TE, Fig. 3B
386 insert). Reciprocally, 232 ORFs were just ribosome-bound in Oocytes, but fully
387 translated in Early embryos, in correlation with an increase in their TE (Fig. 3B, insert).
388 Finally, 3,170 ORFs were translated in our 0-8h AEL samples but did not yield

389 ribosomal FPs in eggs, presumably corresponding to newly-expressed, zygotic-only
390 ORFs (Fig. 3B).

391

392 We hypothesized that a detailed comparison between S2 cells and Late embryos (from
393 where the S2 macrophage-like line was isolated [28]) might reveal the S2 cell-specific
394 translatome profile. Most ORFs seemed translated in both (N=8397), but we observed
395 382 ORFs translated in S2 cells but not in late embryos, whereas 6,669 ORFs
396 appeared translated in late embryos but not in S2 cells (Sup. Fig. 2B). In addition, 540
397 ORFs showed significantly lowered TE in S2 cells and 349 significantly raised TE (Fig.
398 4D). In principle, these differences might reveal a macrophage-like regulatory state, an
399 adaptation to culture, or just a sampling issue (see discussion). Interestingly, we note
400 that the S2 FPs correlate more closely with Early than with Late embryogenesis FPs
401 (Fig. 3A), perhaps reflecting a 'de-differentiation' of the S2 macrophage precursors in
402 cell culture.

403

404 **5) uORFs as translational regulators**

405 We obtain highly reproducible Ribo-Seq signal average for uORFs across biological
406 replicates in the Embryo (*r*=0.95, Sup. Fig. 3A), indicating that most uORFs are bound
407 by ribosomes at specific levels across embryogenesis. Indeed, a majority (72%) of
408 embryo-transcribed uORFs show reproducible ribosome binding during embryogenesis
409 ($RPKM^{FP}$>1 in two replicates, Fig. 2F). However, only 21% of these ribosome-binding
410 events show reliable translation signal, as measured by the binomial framing statistic
411 (Fig. 2F), suggesting that a majority of uORFs in *Drosophila melanogaster* might not
412 have a peptide-productive role.

413

414 Eukaryotic uORFs can act as cis-translational repressors of a canonical main ORF
415 (mORF) located downstream in their transcripts [50]. This regulatory role has been
416 extrapolated to the transcriptome-wide level as the main function of uORFs [51-54].
417 Among canonical protein-coding genes, we saw significant shifts in translational
418 regulation across embryogenesis in hundreds of ORFs (Z-ratio ≥ |1.5|) in both Early-to-
419 Mid (705 ORFs) and Mid-to-Late Embryogenesis (718 ORFs; Figure 4B). However a
420 much smaller number of uORFs exhibited significant variation in translation Z-Ratios

421 (144 and 124 uORFs, respectively; Figure 4C), making it unlikely that translated
422 uORFs underpinned most translational changes in canonical ORFs. However, the cis-
423 regulatory role of uORFs is in principle based on ribosome binding and it is mostly
424 independently of the peptide being produced [50, 55, 56], so it could conceivably be
425 carried out by non-productive ribosomal binding too. Therefore, we extended our
426 analysis to include non-framed, ribosome-bound-only uORFs as well (i.e. $RPKM^{FP}>1$
427 but framing $p>0.01$; Fig. 2A), and this provided a combined pool of 8,346 ribosome-
428 associated uORFs (Fig. 4C).

429

430 Previously, a negative correlation between the number of uORFs within a 5' Leader
431 and the TE of the main ORF was found in Zebrafish genome-wide studies [52] and
432 across vertebrates [53]. We wondered if this was applicable to our 8,346 ribosome-
433 associated uORFs. This would indicate that this negative effect is indeed mediated by
434 uORF ribosomal activity and not a function of other 5´ Leader features [53]. However,
435 we do not see an effect of ribosome-associated uORF number on mORF TE. Instead,
436 we observe a stable TE across mORFs carrying different numbers of uORFs in their
437 5´Leaders (Fig. 5A).

438

439 Nonetheless, it would be possible that a minority of uORFs might have important cis-
440 effects on their mORFs. We looked for this 'uORF onto mORF' regulation by asking if
441 the 1,037 uORFs with significantly changed TE across stages (270 framed and 767
442 ribosome-bound-only, Fig. 4C) also show simultaneous significant TE changes in their
443 cis-associated mORF. We found that not to be the case: 79% (817 of 1,037) of uORFs
444 significantly regulated do not have mORFs with significantly changed TE Z-ratios (≥
445 |1.5|), and vice-versa (Figure 5B). Yet, 12 uORFs correlated negatively with their
446 canonical ORF (i.e. as the uORF TE went up the mORF TE went down, Fig. 5B),
447 suggesting a negative cis-translational regulatory role for these uORFs. Another 25
448 uORFs displayed reduced TE while their respective mORF TEs were upregulated,
449 which is compatible with such a negative regulatory role. However, 183 uORFs showed
450 coordinated changes with their mORF (i.e. uORF and mORF went both up (124) or
451 both down (59) significantly (Fig. 5B), and this positive correlation extended to the TEs
452 of all 8,504 ribosome-associated uORFs and their mORFs (r=0.47 Supp. Fig. 3B). This
453 positive correlation could indicate that uORFs act as either: a) positive translational
454 regulators; b) constant 'brakes' or negative regulators that reduce but do not overcome

455 mORF translation; or finally, c) passive `bystanders´ that are ribosomal-bound by virtue
456 of being present in the 5´Leader of a transcript containing a highly translated mORF. In
457 addition, a combination of these roles is also possible. Interestingly, the total range of
458 TE variation diminishes as the number of ribosome-bound uORFs increases within a
459 given mORF 5' Leader (Fig. 5A), with canonical ORFs that contain no uORFs in their 5´
460 Leaders displaying twice the coefficient of variation as those with one or more
461 ribosome-bound uORFs (Fig. 5A, inset). This suggests a cumulative role for uORFs in
462 the maintenance of mORF translational efficiencies, possibly acting as buffers of
463 variation in mORF translation. A buffering effect could also explain the higher TE
464 averages amongst uORFs than amongst canonical ORFs (Fig. 4B), even though the
465 proportion of ORFs with significantly changed TE was identical (1,037 of 8,504 uORFs:
466 12% vs. 12% canonicals, Fig. 4C - but see also below).

467

468 We observed that 45% percent of canonical ORF 5'-Leaders contain uORFs, a number
469 in line with observations across metazoans [25, 29, 57]. Interestingly, this number is
470 significantly lower in shortCDSs (25%). This could be explained by the difference in 5´
471 Leader average lengths between these two classes (306nt versus 194nt). Indeed, the
472 number of uORFs is positively correlated with the length of the 5´ Leaders across all
473 annotated mRNAs (r=0.668, Fig. 5C), indicating that uORFs tend to accumulate more
474 often in longer 5´-Leaders. The positive correlation between uORF number and
475 5´Leader length supports the idea that uORFs appear randomly in mRNAs [29].

476

477 Whether uORFs act as cis-translational regulators, an important number (1,462, Fig.
478 2F) show framing, i.e. appear to be translated into peptides. On average, $RPKM^{FP}$
479 (Sup. Fig. 3C) and TE (Fig. 4B) was higher in these framed uORFs, and changed more
480 dramatically across stages than on ribosome-bound-only uORFs (note steeper slope in
481 Fig. 4B and higher total percentage of significant translational regulation in Fig. 4C).
482 The regulation of translated uORF expression was not only quantitative, but included
483 qualitative changes as well (i.e. from ribo-bound-only to framed, correlating with an
484 increase in 32nt FPs mapping to the uORF), and transcriptional regulation, altogether
485 producing that 81% of uORFs were translated in a stage-specific manner (Fig. 4A).
486 Thus, regulated expression seemed a feature of uORF translation.

487

488 In the absence of proteomic data to corroborate whether framed uORFs do indeed
489 produce peptides, we checked other bioinformatics markers of coding potential. We
490 observed that translated uORF start codon contexts show significantly better Kozak
491 scores [50] than ribosome-bound-only uORFs (Fig. 5D). Further, the sequence
492 conservation of translated uORFs is intermediate between transcribed uORFs and
493 canonical ORFs (Figure 5E). The amino acid usage produced a similar picture (Figure
494 5F). These observations might suggest that translated uORFs either produce a specific
495 type of peptide, or that translated uORFs are in an evolutionary transition into coding-
496 ness [29]. One end result of such a transition would be to produce a di-cistronic coding
497 gene (uORF plus mORF), and it is interesting that genes annotated as such by
498 FlyBase display similar conservation to canonical ORFs (Fig. 5E), but aminoacid usage
499 intermediate between these and translated uORFs (Fig. 5F).

500

501

502 **DISCUSSION**

503 **Translation vs. peptide function**

504 Protein translation is the most expensive gene activity undertaken by a cell, one order
505 of magnitude more costly than either replication or transcription, and costly enough to
506 have a selective impact [58]. However, it is important to disentangle the act of
507 productive translation from its biological purpose. It is usually expected that translated
508 ORFs must produce a peptide or a protein with canonical function. However, a
509 translated peptide could have other biological activities, [59] or no peptide function at
510 all, such as the act of translation could be the biologically relevant function of the
511 encoding ORF (as in regulatory uORFs, or in canonical ORFs during nonsense-
512 mediated decay) [55, 60]. Ribo-Seq, and other techniques such as proteomics, do not
513 reveal protein and peptide function, only its expression. Protein function is suggested
514 by other studies (such as sequence conservation) and ultimately tested experimentally.

515

516 smORF translation is difficult to ascertain, but has been repeatedly detected to occur at
517 a large scale by Ribo-Seq, including in this study. Different types of studies and
518 analyses yield numbers always in the thousands per genome [45]. These numbers are
519 a small fraction of the smORFs in the genome (about 5%), yet it is large enough (about

520 a 10% of extra coding genes to add to canonical genes) to present a challenge to our
521 understanding of genomes and their function. What can be the biological purpose of
522 this 10% ´crypto-translatome´? Due to their shortness, it is challenging to determine
523 smORF function from their sequence. GO analysis of the longer, and annotated, sCDS
524 smORF class, and analyses of their aminoacid composition [23, 29], suggested roles
525 related to cell membranes and organelles. However, shorter smORFs such as uORFs
526 and those found in lncRNAs (called lncORFs) are not annotated and do not display
527 known protein domains, nor homologies to characterised proteins, that could suggest a
528 peptide function. Their AA composition is intermediate, yet distinct from canonical
529 proteins and random sequences [29]. Nonetheless, important functions for lncORF and
530 uORF peptides have been proven experimentally in a few cases [35, 39, 61-64], as
531 well as a cis-regulatory, 'non-coding' function for uORFs [50] [55, 56].

532

533 **Ribosomal binding versus productive translation**

534 Our results suggest a distinction between ribosomal-binding-only and productive
535 translation. We observed that about 59% of transcribed uORFs show abundant, yet not
536 framed, ribosomal binding. This is unlike canonical ORFs and short CDSs, where only
537 5% and 18% respectively show ribosomal association not corresponding to productive
538 translation. This raises the possibility that either much of ribosomal binding in uORFs
539 serve purposes other than peptide production, or that it reflects 'translational noise'. In
540 view of the energetic and selective costs involved, it is unlikely for the cells to allow for
541 their resources to just be ´lost in translation´. An explanation might be suggested by the
542 maternal translatome.

543

544 **Maternal Translation**

545 The maternal to zygotic transition corroborates that ´ribosomal-binding-only´ without
546 framing is a different state than fully-fledged translation. Our data revealed ORFs
547 ´poised´ for translation between unfertilized eggs and Early embryos. 232 canonical
548 ORFs switch from unproductive ribosome-bound-only to productive translation, while
549 increasing their translational efficiency (Fig. 3B). In other words, the quality of their
550 ribosomal association changes from low affinity (low TE) and not following a particular
551 frame, to higher-affinity, frame-linked productive translation. To produce non-framing,
552 ribosomal binding must happen in overlapping frames, or ´shift´ between frames, or

553 include ´scanning´ 40S subunits, or proofreading ribosomal units involved in NMD and
554 not productively reading codons. Interestingly, we observe one-frame shift of
555 ribosomes at STOP codons (from blue frame “2” to red “0”, Fig. 2B,E), also noticeable
556 in other Ribo-Seq studies [21], which may reflect a conformational change or the
557 ribosome pausing for a time at its A site while disengaging from the mRNA. Under
558 weak translation conditions (indicated by lower $RPKM^{FP}$, Sup. Fig. 3C), in some ORFs
559 elongation may proceed more slowly (perhaps due to a failure to acquire AA-loaded
560 tRNAs or elongation factors). It has been shown that prolonged ribosomal stalling plus
561 queuing can produce both premature termination [65], and use of non-AUG START
562 codons in a different frame [66]. Either effect could produce a frameshifted Ribo-Seq
563 signal towards frames “0” and “1” as we observe, distorting the main ORF signal in
564 frame “2”. In other words, ribosome-bound-only ORFs could be failing to produce
565 enough START to STOP translation, while increasing a frameshifted, noisy ribosomal
566 binding. This effect could be temporary (as in the maternal to zygotic transition) or a
567 constitutive feature of the ORF (as in most uORFs, see below).

568

569 **Function of uORFs**

570 In our previous study [23] we estimated that 34% of Drosophila uORFs were translated
571 in S2 cells, on the basis of a high $RPKM^{FP}$. Interestingly, our improved methods reveal
572 two uORF pools, whose $RPKM^{FP}$ values overlap yet having different averages, and
573 different framing (Sup. Fig. 3C). The percentage of ribosome-bound uORFs in
574 *Drosophila* embryos (71%), is similar to recent estimates [54], but the percentage of
575 actually translated uORFs in our samples is only 12%. The significance of non-
576 productive ribosomal-binding-only in 59% of transcribed uORFs is intriguing. The
577 positive correlation between uORF and mORF with significantly changed TEs, the
578 genome-wide lack of effect of ribosome-bound uORFs on mORF TE, and the
579 correlation between 5´Leader length and number of uORFs all seem to suggest that
580 most uORF are a random and function-less by-product of 5´Leader sequence variation,
581 also suggested by the random-like size distribution of uORFs [29]. Yet, this does not
582 exclude that some uORFs could have a cis-regulatory, or a peptide-producing role.
583 Different mechanisms have been proposed to mediate a cis-regulatory function,
584 including ribosome stalling and disengagement at uORFs, a reduction in translational
585 re-initiation on the downstream mORF AUG codons, and the triggering of mRNA decay
586 [50, 56, 67], all leading to a decrease in mORF translational efficiency. Positive

587 regulatory roles are also possible, with uORFs recruiting ribosomes and thus
588 increasing the likelihood that they reach mORFs via reinitiation, and/or imposing an
589 appropriate ribosomal spacing (reducing ribosome collisions and thus improving mORF
590 TE and fidelity [65, 66, 68]). We propose a combined role, with most uORFs acting as
591 ´translation buffers´ that would recruit ribosomes (a relatively scarce yet valuable
592 cellular resource [68, 69]) and pass them on to their mORF at a constant rate, while
593 discarding ribosomes in excess. A passive, low-level uORF buffer effect at the
594 genome-wide level is consistent with studies in yeasts [70, 71] vertebrates [25] and
595 humans where a genome-wide positive correlation between the TE of individual uORFs
596 and their cognate mORFs is also observed. A function of uORFs as buffers could also
597 generate stalling, queuing, termination and frameshifted starts [66], and explain their
598 frameshifted FP signal.

599

600 A mildly beneficial buffering role could explain the ubiquity of uORFs in otherwise
601 canonical mRNAs. The minority of uORFs showing peptide-productive translation could
602 be a tolerable by-product of this buffering role, which would otherwise be mostly
603 mediated by the much more prevalent uORFs with non-productive ribosomal-binding-
604 only. Interestingly, we observe subtly different Kozak scores, conservation levels and
605 aminoacid usage for translated uORFs, which seem to fit into a stepped continuum
606 ranging from transcribed-only uORFs to canonical ORFs. It is tempting to speculate
607 that, just as ribosomal-bound-only in canonical ORFs may be a developmental
608 transitory state during the maternal to zygotic translation, ribosome-bound-only uORFs
609 may be an evolutionary transitory state poised for a transition to full coding function. In
610 this way, the 5´Leaders of canonical genes would act as ´proto-gene nurseries´ offering
611 a tolerant environment for random ORF creation, while providing coding mRNA
612 features such as poly-A tails, splicing-related and translationally-related RNA
613 processing and stability, allocation to polysomes, and a specific transcriptional pattern.
614 Dicistronic genes could be another step in this transition from inert or regulatory uORF
615 to coding ORF (although dicistronics can also emerge from the fusion of two pre-
616 existing canonical ORFs). Interestingly, short CDSs, which in most aspects studied
617 here (RPKM reproducibility; translation percentage; MassSpec detection; embryonic
618 stage specificity; prevalence of translation regulation; amino-acid usage) are
619 intermediate between translated uORFs and canonical ORFs, do not show
620 intermediate conservation (Fig. 5E). The lower sCDSs conservation may suggest that

621 sCDSs are also involved in another evolutionary process, perhaps another and more
622 dynamic stepped continuum leading to coding-ness via lncRNAs [29, 72].

623

624 **Further developments- single cell translatomics**

625 Despite the improvements to our Poly-Ribo-Seq protocol, and the large number of
626 reads that we have collected, it is possible that not all embryonic translation has been
627 detected in our samples. Lowly abundant or transient mRNAs may not be adequately
628 detected in our 8-hour developmental windows, or may be drowned out by noise and
629 not be able to achieve RPKM >1. This could be especially true of mRNAs expressed in
630 a few cells of the embryo, a situation that must arise often towards the end of
631 embryogenesis, and could be prevalent in differentiated tissues. Our macrophage-like
632 S2 cell line data allows for an approximation to this issue, and the problem of
633 integrating and comparing Ribo-Seq and genomic data from whole embryos and cell
634 lines. Direct comparison between late embryos and S2 cells creates a sampling issue.
635 The embryonic counterparts of S2 cells could be only a few cells amongst many
636 embryonic cell types and some 30,000 cells overall [73, 74]. Thus, the FPs from S2-
637 related ORFs might not be adequately detected in whole embryo extractions (thus
638 yielding a low or no RPKM), whereas the same FPs will be well-detected in the
639 monotypic S2 cell culture. For example, we showed the specific expression (by *in situ*
640 hybridisation and RT-PCR) and function (by observation of mutant phenotypes in
641 embryonic macrophages) of the *hemotin* smORF gene in the *Drosophila* embryo [75].
642 Yet our RNA-Seq and Ribo-Seq data, and modENCODE RNA-Seq would indicate that
643 *hemotin* is neither transcribed nor translated during embryogenesis, while the same
644 data show vigorous expression in S2 cells ($RPKM^{FP}$= 58.6 and framing p=8.15 x $e10^{-5}$).
645 This sampling issue can contrive an artefactual increase of expression and translation
646 (RPKM and TE) of *hemotin* and similar cell-specific transcripts in cell lines [76] when
647 compared with whole embryos or organs. However, a lowering of RPKM or TE in S2
648 cell cultures in 541 ORFs (Fig. 4D) cannot be due to this sampling issue, and must
649 reflect specific translational repression in a S2 ´cell fate´, following either a
650 macrophage-like fate, or a cell-in-culture fate. Thus, Ribo-Seq data from cell lines can
651 reveal genes whose translation declines during specific cell differentiation programmes.
652 Ideally, it would be best to take RiboSeq to this single cell level. The demand for high
653 levels of starting material (in the region of µg) is the largest barrier to single-cell Ribo-
654 Seq, but it could be overcome by the combination of Ribosomal Immunoprecipitation

655 (TRAP) (which requires even more starting material, in the region of mg [77, 78]), with
656 further improvements to the core polysome purification protocol [79].

657

658 **CONCLUSIONS**

659 The likelihood of translation can be determined in Ribo-Seq data by comparing the
660 frequency of reads in frame against a simple binomial probability. Using this metric and
661 a standard RPKM>1 filter for ribosomal binding, almost 16,000 canonical ORFs, more
662 than 300 sCDSs and almost 1,500 uORFs appear translated during the embryonic
663 development of *Drosophila melanogaster*. Translation levels appear more reproducible,
664 yet highly correlated, with transcription, together yielding highly constant expression of
665 canonical ORFs and sCDSs, and higher temporal specificity of uORFs. However, 12-
666 18% ORFs show specific regulation of their translation, including ´poising´ at the
667 maternal to zygotic transition. Finally, a large pool of near 7,000 uORFs also show
668 ribosomal binding without achieving either productive translation, or a significant
669 regulatory input on downstream ORFs. These and other data suggest that, in general,
670 5´-Leaders with uORFs act as ´translation buffers´ at the gene level (helping
671 downstream ORFs to stabilize translation levels), and as ´proto-gene nurseries´ at the
672 genomic level (providing a favourable environment for random ORF creation).

673

674

675 **MATERIALS AND METHODS**

676 **Ribo-Seq and RNA-seq procedures**

677 The RNA-Seq and Poly-Ribo-Seq experiments were conducted as previously
678 described in Aspden et al. [23] with the following modifications. Embryos were flash
679 frozen, turned into powder using a pre-chilled pestle and mortar and homogenized in
680 Lysis Buffer at 4C for 20 min. Pre-clarified lysates were processed for mRNA isolation
681 and fragmentation and polysome separation and digestion as previously described.
682 Digested polysome samples were concentrated by centrifugation and subsequently
683 loaded into a 1M Sucrose Cushion and centrifuged at 70,000g to pellet the
684 monosomes. The 50nt fragmented mRNAs and 28-34 nt footprints were isolated from
685 10% denaturing acrylamide gels and subsequently T4 PNK treated for library
686 preparation.

687

688 The NEBNext Small RNA Library Prep Set for Illumina (NEB, E7300) was used
689 following manufacturers' instructions with small modifications as follows. After the
690 3'Adapters ligation step, ribosome footprints were rRNA depleted using biotinylated
691 DNA fragments and oligos from rRNAs and going through two consecutive rounds of
692 subtractive hybridisation (except for the T 0-8h sample having only one round) using
693 MyOne Streptavidin C1 Dynabeads (Invitrogen). After Ethanol precipitation, samples
694 were processed as NEB guidelines. Libraries were isolated from 10% non-denaturing
695 gels by size selection and their quality and quantity measure by DNA high sensitivity
696 chips (Agilent) with Bioanalyzer and Qubit.

697

698 **Footprint sequence alignment**

699 Ribosomal Footprints were filtered (PHRED quality≥33) and clipped for adapters using
700 the FASTX-Toolkit. The resulting reads were aligned to a FASTA file containing rRNA,
701 tRNA, snoRNA and snRNA from Flybase Release 6.13 annotations using Bowtie,
702 discarding the successful alignments and collecting all unaligned reads. Unaligned
703 reads were then mapped to the FlyBase *D.melanogaster* reference genome (Release
704 6.13) using HISAT2 with default options. Due to the very low frequency of
705 multimapping reads, we retained all genome-aligned reads for further analyses.

706

707 **ORF selection and identification**

708 All annotated coding sequences (CDSs) in FlyBase, r6.13 were retrieved and divided
709 into groups of either longer than or shorter than 303 nucleotides (100 AA). CDSs longer
710 than the cut off were assigned to the canonical ORF category. CDSs shorter or equal
711 to 303 nt. but arising from the same locus as canonicals (short-isoform smORFs, [29])
712 were discarded from our analysis. The remaining short CDSs, which arise from
713 independent genomic loci, were assigned to the short CDS category [29]. We identified
714 uORFs longer than 10 aa with an AUG start codon followed by an in-frame stop codon
715 within the annotated 5′ Leaders of all transcripts annotated as protein-coding in
716 FlyBase r6.13, using the emboss *getorf* program. Next, all ORFs with fully-redundant
717 CDS coordinates were identified, and duplicates were discarded within each class
718 (canonicals, short CDSs and uORFs), Additionally, we only kept uORFs with no

719 overlap with any annotated CDS. uORFs in dicistronic transcripts were then added to
720 the uORF set, based on a full overlap with annotated ORFs contained within the 5'
721 Leader of an immediately downstream annotated CDS.

722

723 **Ribosomal Footprint analysis**

724 Relative Ribosome density ($RPKM^{FP}$) was measured by counting Ribo-Seq reads
725 overlapping each feature using the R package Rsamtols (absolute ribosome density),
726 and scaling this number by feature length and the total number of genome-aligned
727 reads [2]. Total mRNA expression ($RPKM^{RNA}$) was ascertained using the same features
728 and applying the same method on RNA-Seq measurements. Correlations analyses
729 across replicates were measured by Spearman´s rho on ORF RPKM values.

730

731 **Detection of framing and individual translation events**

732 To analyse framing, ribosome footprints (FPs) were first aligned to an artificially
733 constructed transcriptome set, consisting of all ORFs analysed in this study as well as
734 their surrounding regions -18 nucleotides upstream and +15 nucleotides downstream,
735 thus allowing for the alignment of full ribosomal reads spanning the START and STOP
736 codons. The resulting alignments were analysed using the R package RiboSeqR to
737 extract the global framing patterns across ribosome footprint lengths for three pools of
738 RPFs: all pooled embryonic RiboSeq samples, S2 cell samples and pool of mature
739 oocyte and activated eggs RPFs (Kronja et al. [18], cicloheximide^{+}). The main read
740 length and its most overrepresented frame in each sample was then used to evaluate
741 the framing of each ORF in each Ribo-Seq sample. For each ORF, we then compared
742 successful framing events (those matching global framing patterns) with unsuccessful
743 framing events using the binomial test implemented *pbinom()* function in R to measure
744 the probability of obtaining the observed framing (or better), when compared to the
745 expectation by random (1 in 3 probability). For binomial tests of ORF translation,
746 different biological replicates (if available) were merged into a single one. Although the
747 precision of the binomial probability method improves with the amount of evidence (i.e.
748 higher number of reads can achieve very low p-values, or very high probabilities of
749 being translated, Fig. 2C), it can still achieve results with minimal information, which is
750 especially relevant for smORFs as their short size accrues less reads in sequencing
751 experiments when compared to canonical ORFs with the same RPKM range (Fig. 1,

752 2E). Thus, a minimum of 5 reads is required to obtain $p<0.01$ if all reads are in frame,
753 or more than 7 reads if any is out of frame [38].

754

755 **Comparison with other Ribo-Seq and RNA-Seq datasets**

756 We obtained previously-published [18] fastq files containing Cycloheximide^{+} Ribo-Seq
757 reads from mature oocytes and activated eggs from the NCBI Gene Expression
758 Omnibus (accessions SRR1039770 and SRR1039771) as well as the corresponding
759 RNA-Seq experiments (SRR1039764 and SRR1039765), all deposited under
760 accession number GSE52799. The Activated Egg and Oocyte reads were merged to
761 obtain Ribo-Seq and RNA-Seq "Unfertilized Egg" samples, which were analysed
762 employing the same pipeline as our own data. In the case of S2-cell data, we merged
763 our previously-published Ribo-Seq (accessions SRR1548656, SRR1548657,
764 SRR1548658, SRR1548659) and RNA-Seq datasets (SRR1548660- SRR1548661)
765 [23]). We obtained modENCODE RNA-Seq 2-hourly embryonic-staged datasets from
766 FlyBase and consolidated them into our 8-hour-long stages (0h-8h, 8h-16h and 16h-
767 24h AEL).

768

769 **Comparison between RNA-seq and mass spectrometry datasets**

770 We compared the mass spectrometry-based developmental proteome of *Drosophila*
771 *melanogaster* from [26] with our own sequencing data by first averaging the imputed
772 log2 LFQ intensity per protein-group across samples spanning embryonic windows
773 matching those of our timed collections (0h-8h, 8h-16h and 16h-24h AEL). Finally, we
774 normalized all datasets and measured the correlation between the average protein
775 intensities for all identified proteins and our own gene expression measurements using
776 Spearman´s rho.

777

778 **Translation Efficiency and Translational Regulation analysis**

779 Translation efficiency (TE) per ORF was calculated as the fraction of ribosome-density
780 (Poly-Ribo-Seq RPKM) over the total mRNA read density (RNA-Seq RPKM). To detect
781 significant events of translational regulation, we performed Z-ratio calculations of TE

782 variation as per Cheadle et al. 2013 [48]. First, TE values were normalized by
783 calculating the Z-score of each ORF in each sample:

784 $$Zscore_{ORF} = (TE_{ORF} - TE_{sampleMean})/ TE_{sampleSD})$$

785 Second, Z-ratios across contiguous developmental stages e.g. Early and Mid-
786 embryogenesis were calculated for each ORF:

787 $$Zratio = (AverageZScore_{Mid} – Average\ Zscore_{Early)} / (ZScore_{Mid} – Zscore_{Early})_{SD}$$

788 Third, the cut-off of ≥|1.5|was used to identify biologically significant differences [47, 48]
789 between the two contiguous stages.

790

791 **Gene Expression enrichment analysis**

792 We used the unique gene identifiers (FBgn) for genes identified as translationally-
793 regulated across stages (see Translational Regulation analysis) to calculate the
794 embryonic tissue-expression enrichment using the Intermine tool (Flymine), which uses
795 ImaGO terms associated with expression patterns deposited in the BDGP database.
796 The enrichment for each gene set was measured against a background of all
797 *Drosophila* melanogaster genes with expression information, and expressed as a p-
798 value for each significantly-enriched tissue, after the Benjamini-Hochberg multiple-
799 comparisons correction.

800

801 **Conservation analysis**

802 The pre-computed phyloP nucleotide scores for 27-way insect multiple sequence
803 alignments were downloaded from UCSC as bigWig files. We then used the UCSC
804 bigWigAverageOverBed package to compare phyloP scores with BED files containing
805 genome-coordinates for each ORF in our set, obtaining average phyloP scores per
806 ORF.

807

808 **Amino acid composition analysis**

809 We used a previously published own script (*aa_composition.pl*, [23]) to calculate the
810 observed or predicted amino acid compositions of our different sets of ORFs.

811

812 **Kozak-context scoring per ORF**

813 For all canonical ORFs, we extracted the nucleotide composition around-but excluding-
814 the annotated START codon (-5 to 6, excluding positions 1-3). For each position, we
815 then calculated the log odds ratio between observed and background nucleotide
816 frequencies. This provided a scoring table of position-specific nucleotide frequencies
817 from *bona-fide* Kozak contexts with which to score individual ORFs in all classes. The
818 final score per ORF was then obtained by adding the individual position-specific values
819 for all observed nucleotides.

820

821 **DECLARATIONS**

822 **Ethics approval and consent to participate:** Not applicable

823 **Consent for publication:** Not applicable

824 **Availability of data and materials:** Data is retained by the corresponding author and
825 will be made available in a public repository upon publication

826 **Competing interests:** None

827 **Funding:** This work was funded by University of Leeds start-up funds to JA and by
828 Grants from the BBSRC (Ref BB/N001753/1), and the Spanish MINECO (Ref.
829 BFU2016-077793P) to JPC. These bodies had no role in the design of the study or the
830 collection, analysis, and interpretation of data and in writing the manuscript

831 **Authors' contributions:** PP: conception and overall design of the work, bioinformatics
832 analysis and interpretation of Poly-Ribo-Seq, proteomics and genomics data; writing of
833 software and drafting the manuscript. MASM: design of Poly-Ribo-Seq experiments
834 and acquisition of Poly-Ribo-Seq data. JIP: acquisition, bioinformatics analysis and
835 interpretation of Poly-Ribo-Seq data; design and acquisition of tagging data; substantial
836 revision of the manuscript. JLA: interpretation of Poly-Ribo-Seq data; design and
837 acquisition of tagging data, substantial revision of the manuscript. JPC: conception and
838 overall design of the work, interpretation of tagging and Poly-Ribo-Seq data, drafted the
839 manuscript.

Acknowledgements: We thank Rose Phillips for excellent technical assistance, Unum Amin for transfection and confocal microscopy experiments, and Emile Magny and Sarah Newbury for discussions and comments to the manuscript. We also thank Terry Orr-Weaver and Stephen Eichhorn for providing information on the Kronja et al. data.

847 **Figure 1 - Experimental design and quality-control of Poly-Ribo-Seq data. A**
848 *Drosophila melanogaster* embryos were collected in three contiguous 8 hour time-
849 windows spanning the whole of embryogenesis. Two biological replicates were
850 collected for each window, for both Poly-Ribo-Seq and RNA-Seq. pro: procephalon; gb:
851 germ band; df: germ band; mg: presumptive midgut **B** Percentage of 5'UTR, 3'UTR
852 and CDS-aligned reads for Poly-Ribo-Seq, across the annotated transcriptome of
853 *Drosophila melanogaster*. **C** Reproducibility of RPKM values for canonical ORFs by
854 Spearman correlation (*r*) across replicates T (x-axis) and B (y-axis), per stage, for both
855 Poly-Ribo-Seq (top panels) and RNA-Seq (bottom panels). **D** Density of genome-
856 aligned reads for all experiments, both Poly-Ribo-Seq and RNA-Seq across the
857 polycistronic *tarsal-less* (*tal*) transcript.

858

859 **Figure 2 – Detection of Translation in annotated and non-annotated ORFs. A**
860 Sequentially inclusive pipeline used for translation calls. An ORF was considered
861 translated, in each stage, if it showed RPKM > 1 in one of two RNA-Seq replicates,
862 RPKM > 1 in both Poly-Ribo-Seq replicates and a *p*-value < 0.01 in the binomial test for
863 framed-reads. **B** Metagene plot for codon framing of 32-nt. ribosome footprints across
864 all ORFs analysed in this study, showing three-nucleotide periodicity (framing) in the
865 third position of each codon (frame 2, blue). Numbers on top denote the distances
866 between the 5'-end of ribosomal footprints and START/STOP codons. **C** The observed
867 framing for 32-nt. Ribosome footprints is consistent across embryonic stages
868 (combined replicates per stage). **D** Logic of per-ORF analysis of framing probabilities
869 using the binomial test. **E** Density of 32 nt. genome-aligned Poly-Ribo-Seq reads for all
870 experiments, per frame, and corresponding translation probability for the *scl-A* ORF, as
871 measured by the binomial test. **F** Numbers of developmentally-transcribed, ribosome-
872 bound and translated canonical ORFs (annotated ORFs with length > 100 codons),
873 shortCDSs (annotated ORFs ≤ 100 codons) and uORFs.

874

875 **Figure 3 - Comparisons across genome-wide datasets. A** Per-stage Spearman´s
876 correlations across Poly-Ribo-Seq ("FP"), RNA-Seq ("mRNA") and mass-spectrometry
877 datasets ("E"-Early, "M"-Mid, "L"-Late stages correspondence in *Casas-Vila* et al. 2017
878 - see Methods). "Maternal" datasets correspond to concatenated Mature Oocyte and
879 Activated Egg datasets from *Kronja et al. 2014*. S2-cells correspond to the *Aspden et*
880 *al. 2014* dataset plus of novel sequencing. Numbers denote Spearman's *rho*. **B**

881 Numbers and translation fates of Maternal ribosome-bound canonical ORFs in the
882 maternal-to-zygotic transition. “Constant translation” denotes ORFs detected as
883 translated by our pipeline in both Maternal and Early embryo datasets (see Figure 2A).
884 “Poised for translation” denotes ORFs showing Maternal ribosome-binding which only
885 appear as translated in the Early embryo. “Downregulated translation” denotes ORFs
886 with Ribosome-binding and translation in the Maternal dataset, which do not appear in
887 the Early embryo dataset. **C** Overall correlation between average embryo Poly-Ribo-
888 Seq RPKM and average embryo Mass-Spec imputed LFQ intensities (both datasets
889 are log_2-transformed). All detected ORFs are included in this analysis. **D** Differential
890 detectability of canonical ORFs between Poly-Ribo-Seq and Mass-spectrometry
891 techniques. **E** Differential detectability of shortCDSs between Poly-Ribo-Seq and Mass-
892 spectrometry. **F** S2-tagging experiments of 9 uncharacterized short CDSs showing
893 translation in either cortical or reticular-punctate patterns (FLAG antibody: green, F-
894 actin stained with phalloidin: red).

895

896 **Figure 4 - Regulated translation during embryogenesis. A** Number of transcribed,
897 ribosome-bound and translated ORFs per class across *Drosophila melanogaster*
898 embryonic development, showing progressive specificity in ON/OFF translation
899 patterns. **B** Changes in average translation efficiency, per ORFs class, across
900 analysed developmental time-points. **C** Z-score ratio analysis of translation efficiencies
901 (TE) during successive developmental stages pinpointing ORFs subjected to significant
902 developmental modulations of translation across ORF classes. **D** Number of
903 developmentally-regulated canonical ORFs subjected to downregulation (bottom, Z-
904 ratio ≤ -1.5) or upregulation (top, Z-ratio ≥ 1.5), per developmental transition. **E-F**
905 Significantly-enriched gene expression patterns in the translationally-upregulated
906 canonical genes the Early-to-Mid and Maternal-to-Early developmental transitions
907 respectively (*p-value* of BGDP expression annotations after Benjamini-Hochberg
908 procedure).

909

910 **Figure 5 - uORFs as translational regulators. A** Variation in canonical TE levels as a
911 function of the number of cistronic uORFs showing ribosomal binding. **B** Comparison
912 of translational changes between uORFs and cognate canonical ORFs across
913 developmental stages. Numbers denote the number of Z-ratio significant uORF-mORF
914 pairs in each quadrant. Vertical arrows: uORF translational regulation sign. Horizontal

915 arrows: cognate canonical translational regulation sign. Green arrows: Upregulated;
916 Red arrows: Downregulated. **C** Correlation between the number of uORFs per mRNA
917 and 5'UTR length (Spearman's r = 0.7). **D** Comparison of Kozak-context scores
918 between distinct ORF classes and uORF subclasses. Graphs display 10-90 percentile
919 range. Mann-Whitney $p < 0.001$. **E** Comparison of phyloP conservation scores (27-way
920 alignments) between distinct ORF classes and uORF subclasses. Graphs display 10-
921 90 percentile range. Mann-Whitney $p < 0.01$. **F** Correlations in amino acid usage
922 across distinct uORF subclasses, shortCDSs and canonical ORFs; values denote
923 pairwise Spearman's r. **G-H**: FLAG antibody: green, F-actin stained with phalloidin: red.
924 Scale bar denotes 5 micrometres. **G** uORF evolution model. 5'-leaders in blue. t:
925 evolutionary time. Red lines denote ribosomal binding signal; white boxes correspond
926 to novel uORFs; salmon boxes correspond to ribosomal-bound uORFs; red boxes
927 correspond to translated uORFs.

928

929 **Supplementary Figure 1 – RiboSeq read length distribution and transcriptome-**
930 **wide framing. A** Distribution of Poly-Ribo-Seq genome-mapped ribosome-protected
931 fragment (RPFs) lengths; codon-framing of predominant lengths shown in inset. **B**
932 negative control: of codon-framing in genome-mapped RNA-Seq reads fails to show
933 dominance in any of the frames.

934

935 **Supplementary Figure 2 –A** Detectability of mRNA translation across techniques.
936 Total number of sCDSs detected by Mass-spectrometry, Poly-Ribo-Seq and FLAG-
937 Tagging (Total N=44). **B** S2-cell canonical translation. Comparison of translated
938 canonical ORFs in S2 cells and the late embryo stages from which these cell culture
939 primarily originated [28].

940

941 **Supplementary Figure 3 – uORF A** Comparison of uORF translation from T and B
942 replicas (in $\log_2$RPKM) show a very high correlation (Spearman's r=0.9466). **B**
943 Histogram of individual Spearman's correlation r between TE of ribosome-bound
944 uORFs and their corresponding mORF across embryogenesis. The overall positive
945 correlation has a median of r=0.477 (binning=0.1 windows). **C**. Distribution of
946 $\log_2(\mathrm{RPKM}^{FP})$ values of ribosome-bound-only uORFs (red) and translated (framed)
947 uORFs (green).

948

949 **Supplementary Table 1 – Sequencing statistics for novel RiboSeq and RNA-Seq**
950 **datasets reported in this study.** Number of raw reads per replica per stage and
951 percentage of remaining reads after ribosomal RNA subtraction ("norRNA") and
952 genome-mapping (see methods).

953

954 **Supplementary Table 2 – Heatmap of correlations across RNA-Seq and RiboSeq**
955 **samples and replicates** Spearman correlation values (*r*) for canonical ORF RPKMs
956 across all RNA-Seq and RiboSeq replicates analysed in this study. All correlations
957 were significant ($p<0.0001$). T and B denote the two biological replicates used in this
958 study. Early, Mid and Late denote Embryonic windows 0-8h, 8-16h and 16-24h,
959 respectively. "Flybase" RNA-Seq refers to modENCODE data (Graveley et al. 2011).
960 S2 refers to S2-cell datasets (see methods). Unfertilized Egg corresponds to
961 previously-published data (Kronja et al. 2014).

962

963 **Supplementary Tables 3 - Gene expression, regulation and conservation**
964 **statistics for distinct small ORFs in polycistronic mRNAs. A** The *tal* polycistronic
965 mRNA shows marked quantitative differences in translation efficiency, probability and
966 regulation, as well as conservation, across its four small ORFs. **B** The *scl* locus
967 contains two cistronic small ORFs that show similar but quantitatively different levels of
968 translation and conservation.

969

970 **Supplementary Table 4 – Number of Transcribed, Ribosome-Bound and**
971 **Translated ORFs by Class and embryonic stage.** Different ORF classes show
972 distinct amounts of transcription, ribosome binding and translation across stages.

973

974

975 1. Ingolia NT, Lareau LF, Weissman JS: **Ribosome profiling of mouse embryonic stem**
976 **cells reveals the complexity and dynamics of Mammalian proteomes.** *Cell* 2011,
977 **147:**789-802.

978 2. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: **Genome-wide analysis in**
979 **vivo of translation with nucleotide resolution using ribosome profiling.** *Science*
980 2009, **324:**218-223.

981 3. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills
982 MR, Weissman JS: **Ribosome profiling reveals pervasive translation outside of**
983 **annotated protein-coding genes.** *Cell Rep* 2014, **8:**1365-1379.

984 4. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, Landthaler M,
985 Obermayer B, Ohler U: **Detecting actively translated open reading frames in**
986 **ribosome profiling data.** *Nat Methods* 2016, **13:**165-170.

987 5. Mumtaz MA, Couso JP: **Ribosomal profiling adds new coding sequences to the**
988 **proteome.** *Biochem Soc Trans* 2015, **43:**1271-1276.

989 6. Jang C, Lahens NF, Hogenesch JB, Sehgal A: **Ribosome profiling reveals an**
990 **important role for translational control in circadian gene expression.** *Genome Res*
991 2015, **25:**1836-1847.

992 7. Spriggs KA, Bushell M, Willis AE: **Translational regulation of gene expression**
993 **during conditions of cell stress.** *Mol Cell* 2010, **40:**228-237.

994 8. Sendoel A, Dunn JG, Rodriguez EH, Naik S, Gomez NC, Hurwitz B, Levorse J, Dill
995 BD, Schramek D, Molina H, et al: **Translation from unconventional 5′ start sites**
996 **drives tumour initiation.** *Nature* 2017, **541:**494-499.

997 9. Hinnebusch AG, Ivanov IP, Sonenberg N: **Translational control by 5'-untranslated**
998 **regions of eukaryotic mRNAs.** *Science* 2016, **352:**1413-1416.

999 10. Stadler M, Fire A: **Conserved translatome remodeling in nematode species**
1000 **executing a shared developmental transition.** *PLoS genetics* 2013, **9:**e1003739.

1001 11. Gilbert SF: *Developmental Biology.* Sinauer Associates; 1997.

1002 12. Zalokar M: **Autoradiographic study of protein and RNA formation during early**
1003 **development of Drosophila eggs.** *Dev Biol* 1976, **49:**425-437.

1004 13. Qin X, Ahn S, Speed TP, Rubin GM: **Global analyses of mRNA translational control**
1005 **during early Drosophila embryogenesis.** *Genome Biol* 2007, **8:**R63.

1006 14. Yoshikane N, Nakamura N, Ueda R, Ueno N, Yamanaka S, Nakamura M: **Drosophila**
1007 **NAT1, a homolog of the vertebrate translational regulator NAT1/DAP5/p97, is**
1008 **required for embryonic germband extension and metamorphosis.** *Dev Growth*
1009 *Differ* 2007, **49:**623-634.

1010 15. Bate M, Martinez-Arias, A. (Ed.). **The development of Drosophila melanogaster**:
1011 Cold Spring Harbour Laboratory Press; 1993.

1012 16. Ray RP, Arora K, Nusslein-Volhard C, Gelbart WM: **The control of cell fate along the**
1013 **dorsal-ventral axis of the Drosophila embryo.** *Development* 1991, **113:**35-54.

1014 17. Stjohnston D, Nussleinvolhard C: **The Origin of Pattern and Polarity in the**
1015 **Drosophila Embryo.** *Cell* 1992, **68:**201-219.

1016 18. Kronja I, Yuan B, Eichhorn SW, Dzeyk K, Krijgsveld J, Bartel DP, Orr-Weaver TL:
1017 **Widespread changes in the posttranscriptional landscape at the Drosophila oocyte-**
1018 **to-embryo transition.** *Cell Rep* 2014, **7:**1495-1508.

1019 19. Tadros W, Lipshitz HD: **The maternal-to-zygotic transition: a play in two acts.**
1020 *Development* 2009, **136:**3033-3042.

1021 20. Yartseva V, Giraldez AJ: **The Maternal-to-Zygotic Transition During Vertebrate**
1022 **Development: A Model for Reprogramming.** *Curr Top Dev Biol* 2015, **113:**191-232.

1023 21. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES,
1024 Vejnar CE, Lee MT, Rajewsky N, Walther TC, Giraldez AJ: **Identification of small**
1025 **ORFs in vertebrates using ribosome footprinting and evolutionary conservation.** In
1026 *EMBO J*, vol. 33, 2014/04/08 edition. pp. 981-993; 2014:981-993.

1027 22. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G,
1028 Rajewsky N, Kempa S, Selbach M, Obermayer B: **Extensive identification and**
1029 **analysis of conserved small ORFs in animals.** *Genome Biol* 2015, **16:**179.

1030 23. Aspden JL, Eyre-Walker YC, Philips RJ, Brocard M, Amin U, Couso JP: **Extensive**
1031 **translation of small Open Reading Frames revealed by Poly-Ribo-Seq** *eLife* 2014,
1032 **3:**e03528.

1033 24. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES: **Ribosome profiling**
1034 **provides evidence that large noncoding RNAs do not encode proteins.** *Cell* 2013,
1035 **154:**240-251.

1036 25. Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, Valen E: **Ribosome profiling**
1037 **reveals resemblance between long non-coding RNAs and 5' leaders of coding**
1038 **RNAs.** *Development* 2013, **140:**2828-2834.

1039 26. Casas-Vila N, Bluhm A, Sayols S, Dinges N, Dejung M, Altenhein T, Kappei D,
1040 Altenhein B, Roignant JY, Butter F: **The developmental proteome of Drosophila**
1041 **melanogaster.** *Genome Res* 2017, **27:**1273-1285.

1042 27. Heyer Erin E, Moore Melissa J: **Redefining the Translational Status of 80S**
1043 **Monosomes.** *Cell* 2016, **164:**757-769.

1044 28. Schneider I: **Cell lines derived from late embryonic stages of Drosophila**
1045 **melanogaster.** *J Embryol Exp Morphol* 1972, **27:**353-365.

1046 29. Couso JP, Patraquim P: **Classification and function of small open reading frames.**
1047 *Nat Rev Mol Cell Biol* 2017, **18:**575-589.

1048 30. Martinez-Arias A: **Development and patterning of the larval epidermis of**
1049 **Drosophila.** In *The development of Drosophila melanogaster.* Edited by Bate CM,
1050 Martinez Arias A: Cold Spring Harbour Laboratory Press; 1993: 517-608

1051 31. Skaer H: **The Alimentary Canal.** In *The development of Drosophila melanogaster.*
1052 Edited by Bate.M, Martinez Arias.A: Cold Spring Harbor Laboratory Press; 1993: 941-
1053 1012

1054 32. Miguel-Aliaga I, Jasper H, Lemaitre B: **Anatomy and Physiology of the Digestive**
1055 **Tract of Drosophila melanogaster.** *Genetics* 2018, **210:**357-396.

1056 33. Bate M: **The embryonic development of larval muscles in Drosophila.** *Development*
1057 1990, **110:**791-804.

1058 34. Landgraf M, Jeffrey V, Fujioka M, Jaynes JB, Bate M: **Embryonic origins of a motor**
1059 **system: motor dendrites form a myotopic map in Drosophila.** *PLoS biology* 2003,
1060 **1:**e41.

1061 35. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP: **Peptides encoded by short**
1062 **ORFs control development and define a new eukaryotic gene family.** *PLoS Biology*
1063 2007, **5:**1052-1062.

1064 36. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS: **Ribosome profiling reveals**
1065 **pervasive and regulated stop codon readthrough in Drosophila melanogaster.** *eLife*
1066 2013, **2:**e01179.

1067 37. Calviello L, Ohler U: **Beyond Read-Counts: Ribo-seq Data Analysis to Understand**
1068 **the Functions of the Transcriptome.** *Trends Genet* 2017, **33:**728-744.

1069 38. Zar JH: *Biostatistical analysis.* 2019.

1070 39. Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, Couso JP:
1071 **Conserved regulation of cardiac calcium uptake by peptides encoded in small open**
1072 **reading frames.** *Science* 2013, **341:**1116-1120.

1073 40. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG,
1074 van Baren MJ, Boley N, Booth BW, et al: **The developmental transcriptome of**
1075 **Drosophila melanogaster.** *Nature* 2011, **471:**473-479.

1076 41. Loose CR, Langer RS, Stephanopoulos GN: **Optimization of protein fusion partner**
1077 **length for maximizing in vitro translation of peptides.** *Biotechnol Prog* 2007,
1078 **23:**444-451.

1079 42. Jenssen H, Aspmo SI: **Serum stability of peptides.** *Methods Mol Biol* 2008, **494:**177-
1080 186.

1081 43. Amin U: **Cellular characterisation of small Open Reading Frame function in**
1082 **Drosophila melanogaster.** *Doctoral Thesis.* University of Sussex, Biology; 2016.

1083 44. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch
1084 EW, Panse C, de Lichtenberg U, Rinner O, et al: **A high-quality catalog of the**
1085 **Drosophila melanogaster proteome.** *Nat Biotechnol* 2007, **25:**576-583.

1086 45. Pueyo JI, Magny EG, Couso JP: **New peptides under the s(ORF)ace of the genome.**
1087 *Trends in Biochemical Sciences* 2016, **41:**665-678.

1088 46. Zhong Y, Karaletsos T, Drewe P, Sreedharan VT, Kuo D, Singh K, Wendel HG, Ratsch
1089 G: **RiboDiff: detecting changes of mRNA translation efficiency from ribosome**
1090 **footprints.** *Bioinformatics* 2017, **33:**139-141.

1091 47. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet*
1092 2002, **32 Suppl:**496-501.

1093 48. Cheadle C, Vawter MP, Freed WJ, Becker KG: **Analysis of microarray data using Z**
1094 **score transformation.** *J Mol Diagn* 2003, **5:**73-81.

1095 49. Thomsen S, Anders S, Janga S, Huber W, Alonso C: **Genome-wide analysis of mRNA**
1096 **decay patterns during early Drosophila development.** *Genome Biology* 2010, **11:**1-
1097 27.

1098 50. Kozak M: **Regulation of translation via mRNA structure in prokaryotes and**
1099 **eukaryotes.** *Gene* 2005, **361:**13-37.

1100 51. Calvo SE, Pagliarini DJ, Mootha VK: **Upstream open reading frames cause**
1101 **widespread reduction of protein expression and are polymorphic among humans.**
1102 *Proc Natl Acad Sci U S A* 2009, **106:**7507-7512.

1103 52. Johnstone TG, Bazzini AA, Giraldez AJ: **Upstream ORFs are prevalent translational**
1104 **repressors in vertebrates.** *Embo J* 2016, **35:**706-723.

1105 53. Chew GL, Pauli A, Schier AF: **Conservation of uORF repressiveness and sequence**
1106 **features in mouse, human and zebrafish.** *Nat Commun* 2016, **7:**11663.

1107 54. Zhang H, Dou S, He F, Luo J, Wei L, Lu J: **Genome-wide maps of ribosomal**
1108 **occupancy provide insights into adaptive evolution and regulatory roles of uORFs**
1109 **during Drosophila development.** *PLoS Biol* 2018, **16:**e2003903.

1110 55. Somers J, Poyry T, Willis AE: **A perspective on mammalian upstream open reading**
1111 **frame function.** *Int J Biochem Cell Biol* 2013, **45:**1690-1700.

1112 56. Andrews SJ, Rothnagel JA: **Emerging evidence for functional peptides encoded by**
1113 **short open reading frames.** *Nat Rev Genet* 2014, **15:**193–204.

1114 57. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A,
1115 Rinn JL, Regev A, Schier AF: **Systematic identification of long noncoding RNAs**
1116 **expressed during zebrafish embryogenesis.** *Genome research* 2012, **22:**577-591.

1117 58. Lynch M, Marinov GK: **The bioenergetic costs of a gene.** *Proceedings of the National*
1118 *Academy of Sciences* 2015, **112:**15690.

1119 59. Starck SR, Tsai JC, Chen K, Shodiya M, Wang L, Yahiro K, Martins-Green M, Shastri
1120 N, Walter P: **Translation from the 5' untranslated region shapes the integrated**
1121 **stress response.** *Science (New York, NY)* 2016, **351:**aad3867-aad3867.

1122 60. Kervestin S, Jacobson A: **NMD: a multifaceted response to premature translational**
1123 **termination.** *Nat Rev Mol Cell Biol* 2012, **13:**700-712.

1124 61. Combier JP, de Billy F, Gamas P, Niebel A, Rivas S: **Trans-regulation of the**
1125 **expression of the transcription factor MtHAP2-1 by a uORF controls root nodule**
1126 **development.** *Genes & Development* 2008, **22:**1549-1559.

1127 62. Hanyu-Nakamura K, Sonobe-Nojima H, Tanigawa A, Lasko P, Nakamura A:
1128 **Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial**
1129 **germ cells.** *Nature* 2008, **451:**730-733.

1130 63. Pauli A, Norris ML, Valen E, Chew GL, Gagnon JA, Zimmerman S, Mitchell A, Ma J,
1131 Dubrulle J, Reyon D, et al: **Toddler: an embryonic signal that promotes cell**
1132 **movement via Apelin receptors.** *Science* 2014, **343:**1248636.

1133 64. Slavoff SA, Heo J, Budnik BA, Hanakahi LA, Saghatelian A: **A human short open**
1134 **reading frame (sORF)-encoded polypeptide that stimulates DNA end joining.** *J*
1135 *Biol Chem* 2014, **289:**10950-10957.

1136 65. Ferrin MA, Subramaniam AR: **Kinetic modeling predicts a stimulatory role for**
1137 **ribosome collisions at elongation stall sites in bacteria.** *eLife* 2017, **6:**e23629.

1138 66. Ivanov IP, Shin BS, Loughran G, Tzani I, Young-Baird SK, Cao C, Atkins JF, Dever
1139 TE: **Polyamine Control of Translation Elongation Regulates Start Site Selection on**
1140 **Antizyme Inhibitor mRNA via Ribosome Queuing.** *Mol Cell* 2018, **70:**254-264.e256.

1141 67. Barbosa C, Peixeiro I, Romao L: **Gene expression regulation by upstream open**
1142 **reading frames and human disease.** *PLoS Genet* 2013, **9:**e1003529.

1143 68. Tarrant D, von der Haar T: **Synonymous codons, ribosome speed, and eukaryotic**
1144 **gene expression regulation.** *Cellular and Molecular Life Sciences* 2014, **71:**4195-
1145 4206.

1146 69. Warner JR: **The economics of ribosome biosynthesis in yeast.** *Trends Biochem Sci*
1147 1999, **24:**437-440.

1148 70. Duncan CD, Mata J: **The translational landscape of fission-yeast meiosis and**
1149 **sporulation.** *Nat Struct Mol Biol* 2014, **21:**641-647.

1150 71. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS: **High-**
1151 **resolution view of the yeast meiotic program revealed by ribosome profiling.**
1152 *Science* 2012, **335:**552-557.

1153 72. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM: **Long non-coding RNAs as a**
1154 **source of new peptides.** *Elife* 2014, **3:**e03523.

1155 73. Hartenstein V, Campos-Ortega JA: **Fate-mapping in wild-type Drosophila**
1156 **melanogaster. I. The spatio-temporal pattern of embryonic cell divisions.** *Roux*
1157 *Arch dev Biol* 1985, **194:**181-195.

1158 74. Williams MJ: **Drosophila hemopoiesis and cellular immunity.** *J Immunol* 2007,
1159 **178:**4711-4716.

1160 75. Pueyo JI, Magny EG, Sampson CJ, Amin U, Evans IR, Bishop SA, Couso JP:
1161 **Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved**
1162 **across metazoans.** *PloS Biology* 2016, **14:**e1002395.

1163 76. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A,
1164 Lagarde J, Lin W, Schlesinger F, et al: **Landscape of transcription in human cells.**
1165 *Nature* 2012, **489:**101-108.

1166 77. Heiman M, Kulicke R, Fenster RJ, Greengard P, Heintz N: **Cell type-specific mRNA**
1167 **purification by translating ribosome affinity purification (TRAP).** *Nat Protoc* 2014,
1168 **9:**1282-1291.

1169 78. Chen X, Dickman D: **Development of a tissue-specific ribosome profiling approach**
1170 **in Drosophila enables genome-wide evaluation of translational adaptations.** *PLoS*
1171 *Genet* 2017, **13:**e1007117.

1172 79. Liang S, Bellato HM, Lorent J, Lupinacci FCS, Oertlin C, van Hoef V, Andrade VP,
1173 Roffe M, Masvidal L, Hajj GNM, Larsson O: **Polysome-profiling in small tissue**
1174 **samples.** *Nucleic Acids Res* 2018, **46:**e3.

1175

Figure 1

Method and basic data

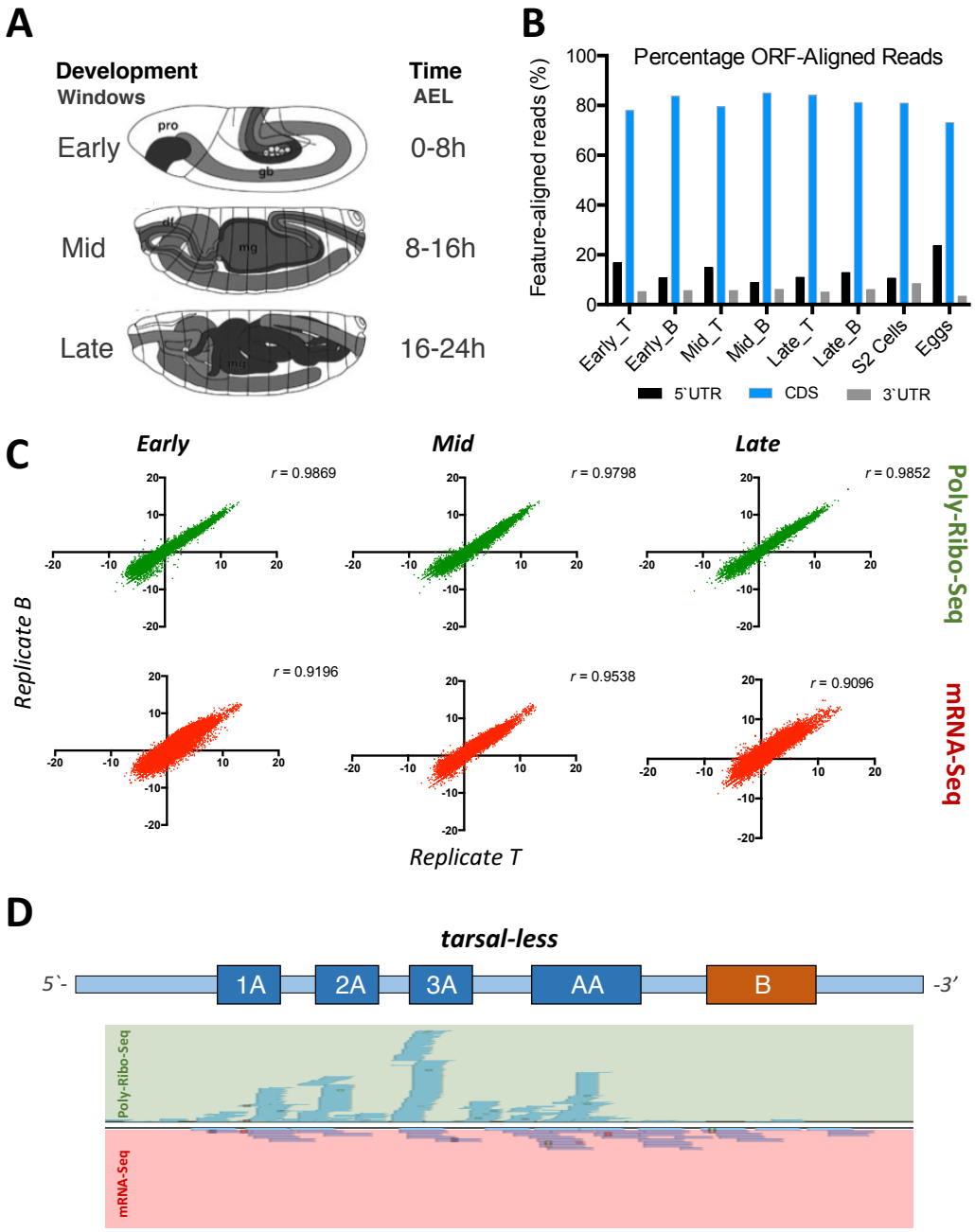


Figure 2

Detecting Translation

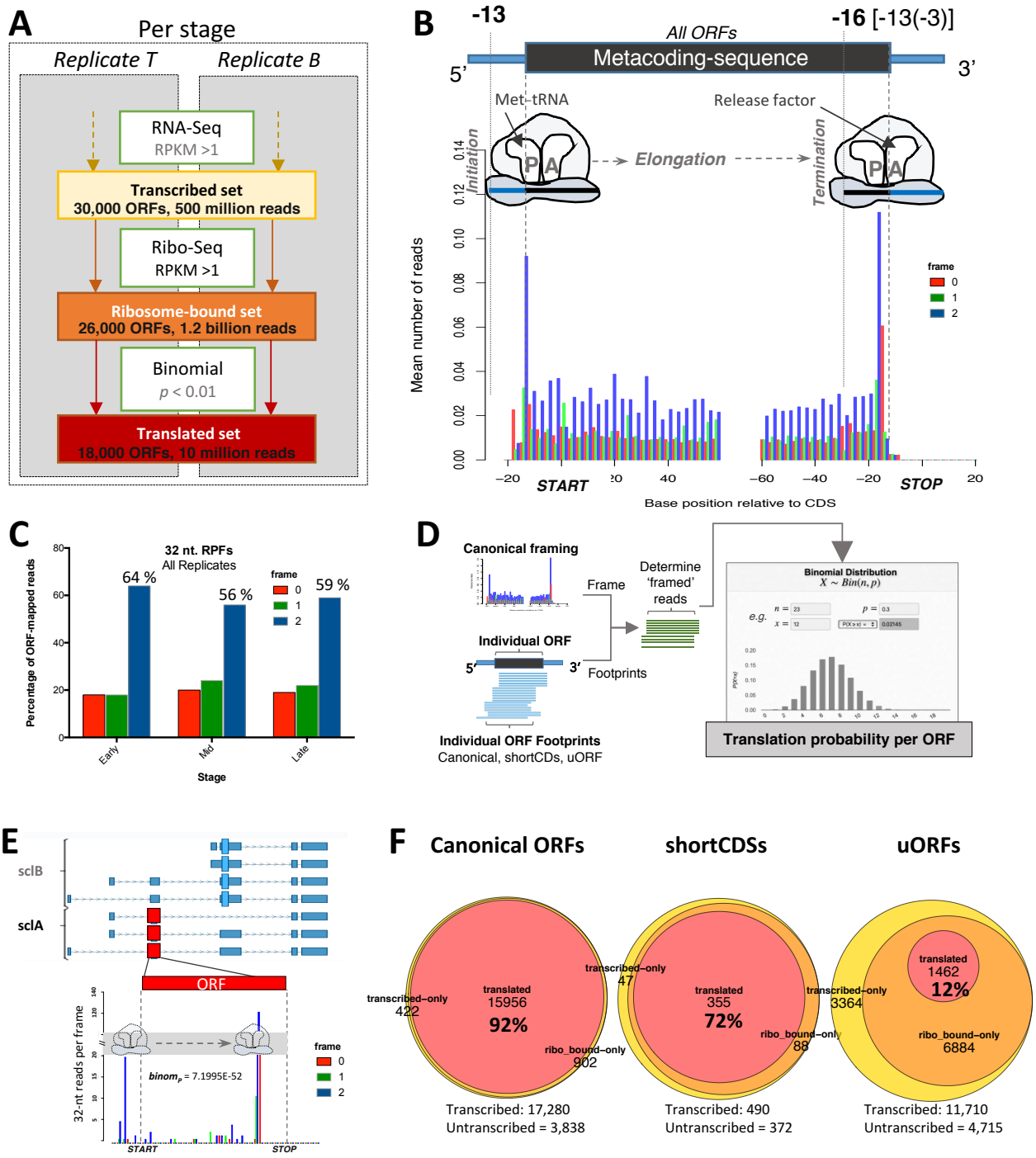


Figure 3

Comparison among genomic data

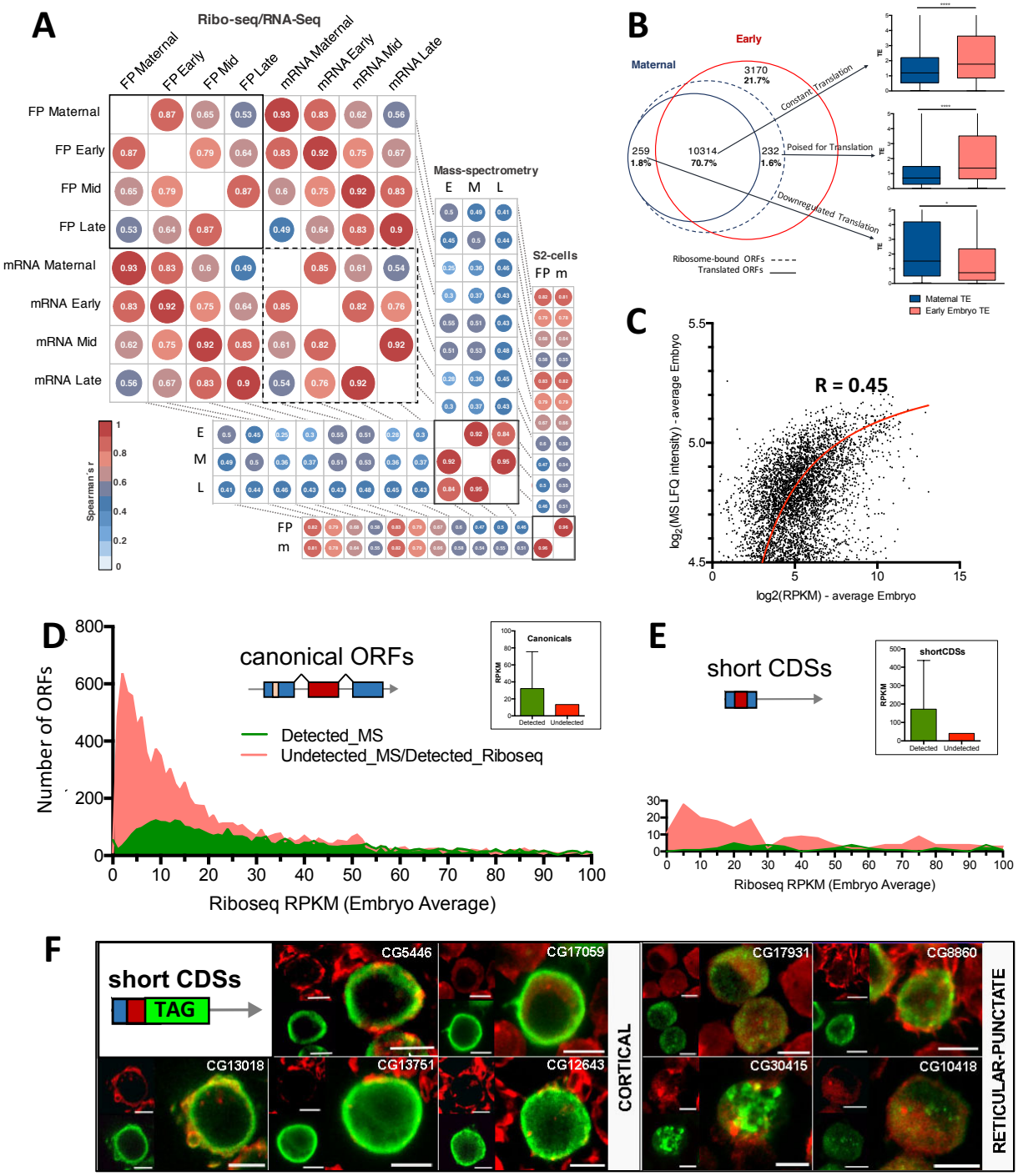
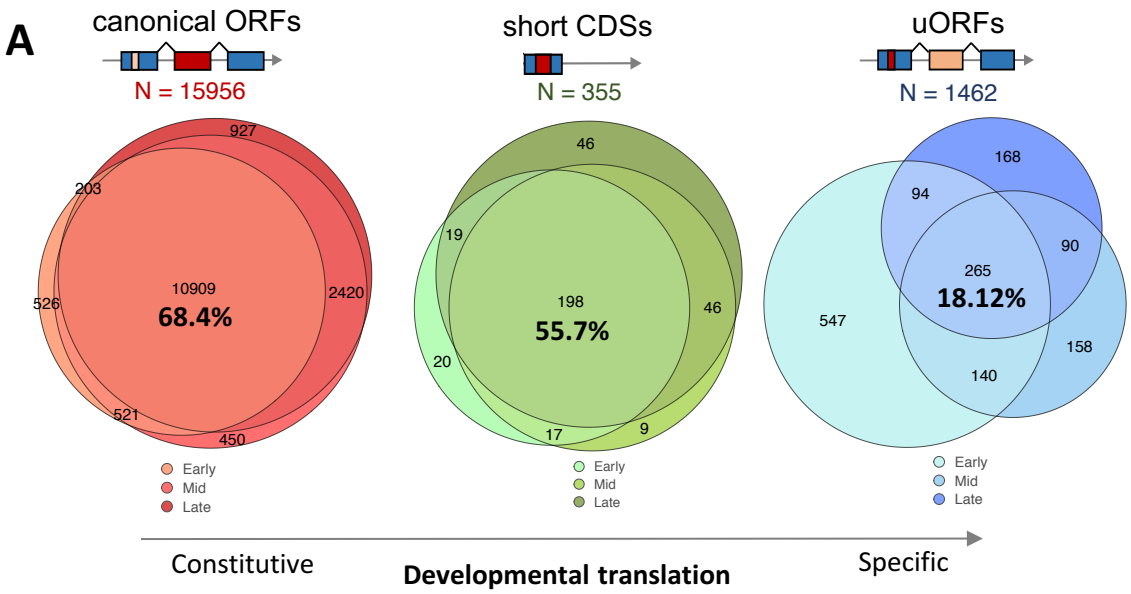
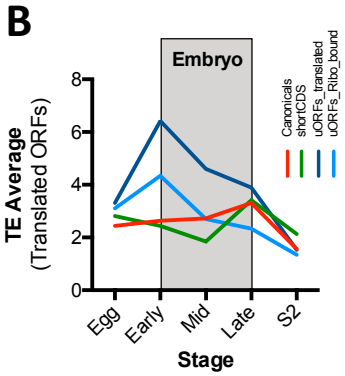


Figure 4

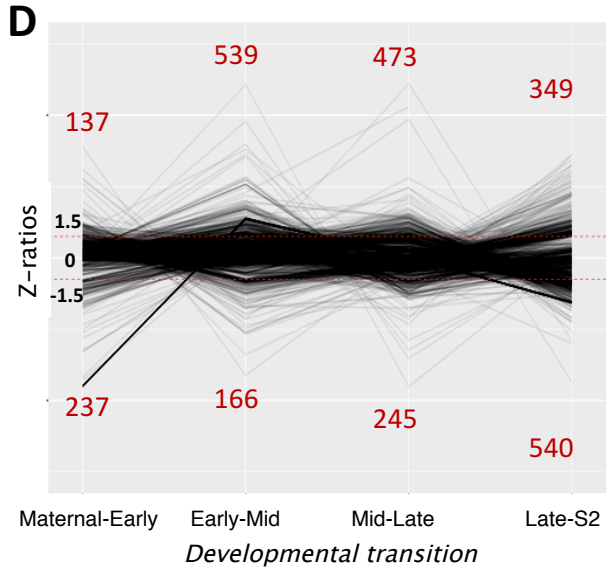
Regulated translation during embryogenesis





C

			Translationally-regulated			
Class	**Translated**	**Maternal-to-Early**	**Early-to-Mid**	**Mid-to-Late**	**Late-to-S2**	**Total**
Canonicals	15956	374	705	718	889	1899 (**12%**)
Short CDSs	355	6	20	29	20	57 (**16%**)
$uORFs_{translated}$	1462	78	144	124	59	263 (**18%**)
$uORFs_{ribo\text{-}bound}$	6884	216	346	328	185	755 (**11%**)



E

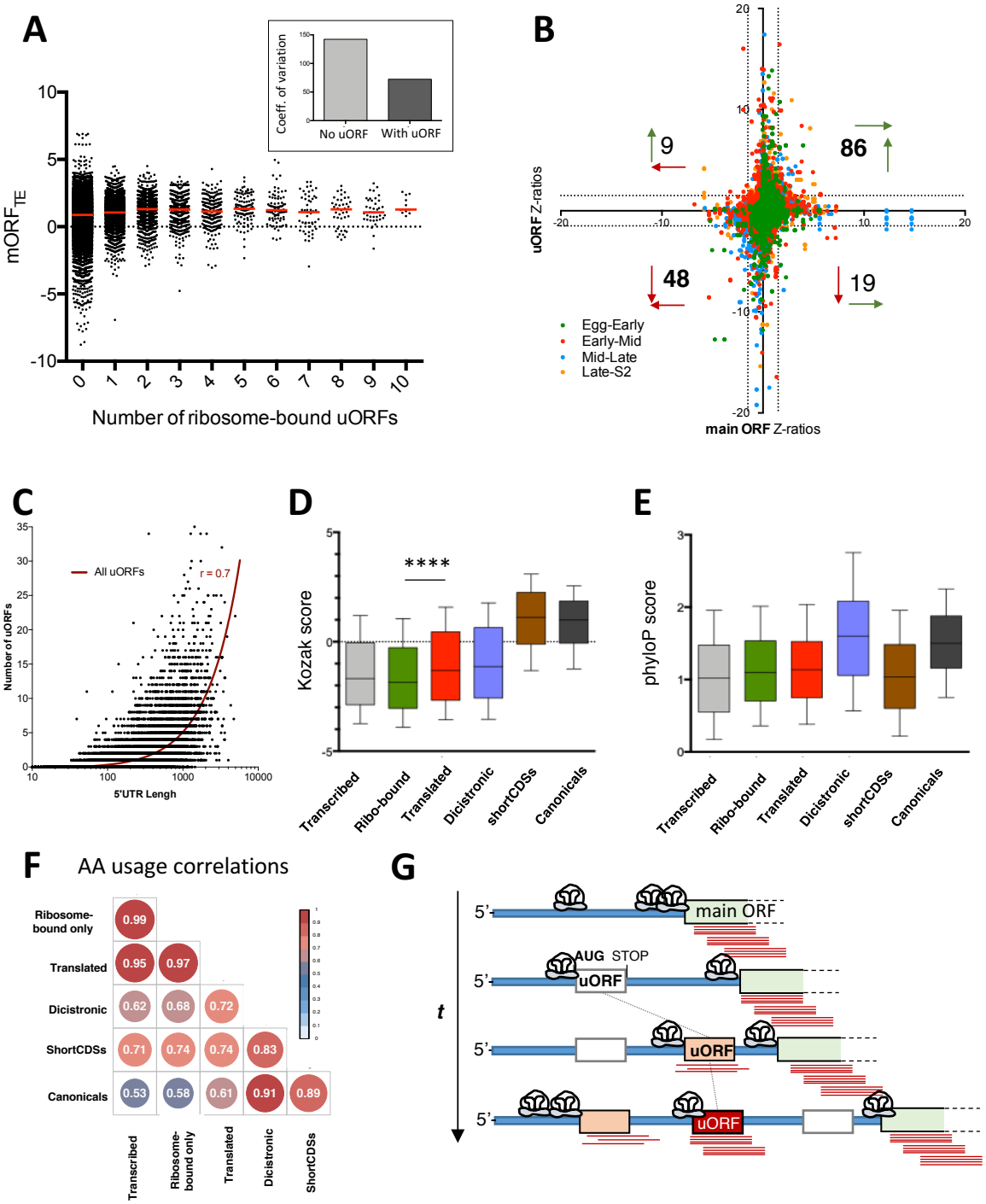
228 genes (539 ORFs) - Up Early-to-Mid

Terms	p-Value	Matches
ventral nerve cord	6.345764e-4	52
embryonic dorsal epidermis	0.001006	35
embryonic brain	0.002130	49
embryonic epipharynx	0.002235	20
embryonic hypopharynx	0.003883	20
embryonic foregut	0.005816	22
embryonic esophagus	0.005891	14
embryonic ventral epidermis	0.008728	29
embryonic/larval tracheal system	0.011728	17

F

100 genes (161 ORFs)- Up Maternal-to-Early

Terms	p-Value	Matches
ubiquitous	0.003625	35
maternal	0.038710	49

Figure 5

uORFs as translational regulators