Molecular and functional characterization of the *Drosophila melanogaster* conserved smORFome

Authors:

Justin A. Bosch[1,6], Nathan Keith[2,3,6], Felipe Escobedo[1], William F. Fisher[2], James Thai LaGraff[1], Jorden Rabasco[1], Kenneth H. Wan[2], Richard Weiszmann[2], Yanhui Hu[1], Shu Kondo[4], James B. Brown[2,3,7], Norbert Perrimon[1,5,7], Susan E. Celniker[2,7,8]

[1]Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA

[2]Division of Biological Systems and Engineering, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[3]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[4]Laboratory of Invertebrate Genetics, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

[5]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

[6]These authors contributed equally

[7]These authors jointly supervised this work

[8]Lead Contact

Correspondence: jabosch@hms.harvard.edu (J.B.), rnkeith@lbl.gov (N.K.), jbbrown@lbl.gov (J.B.B.), perrimon@genetics.med.harvard.edu (N.P.) secelniker@lbl.gov (S.C.)

Summary:

Short polypeptides encoded by small open reading frames (smORFs) are ubiquitously found in eukaryotic genomes and are important regulators of physiology, development, and mitochondrial processes. Here, we focus on a subset of 194 smORFs that are evolutionarily conserved between *Drosophila melanogaster* and humans. Many of these smORFs are conserved broadly in the bilaterian lineage, with ~60 conserved in plants. Within these conserved smORFs, we observed remarkably heterogenous spatial and temporal expression patterns – indicating wide-spread tissue-specific and stage-specific mitochondrial architectures. In addition, an analysis of annotated functional domains revealed a predicted enrichment of smORF polypeptides localizing to mitochondria. We conducted an embryonic ribosome profiling experiment and applied a machine learning approach, finding support for translation of 82 of these smORFs during embryogenesis. We further embarked on functional characterization using CRISPR knockout/activation, RNAi knockdown, and cDNA overexpression, revealing diverse phenotypes. This study underscores the importance of identifying smORF function in disease and phenotypic diversity.

47 **Introduction:**

48 Genome annotations have often overlooked proteins with less than 100 amino acids,
49 although many have been shown to play important roles in development and physiology
50 and are pervasive across the Tree of Life (Plaza et al., 2017; Couso and Patraquim,
51 2017). While some small proteins are cleavage products of longer proteins, many others
52 are encoded in the genome by small open reading frames (smORF genes; ≤ 100 amino
53 acids). Strikingly, among the currently 681 annotated human smORF genes, 30%
54 include disease-associated variants mapped within exons, compared to 15% of human
55 genes in general, according to the NHGRI GWAS catalog (Welter et al., 2014). Although
56 this two-fold enrichment is partially due to the fact that only the most highly conserved
57 smORF genes have been annotated due to the strict length cutoffs used during first-
58 generation genome annotations (Basrai et al., 1997; Olexiouk et al., 2018), these
59 estimates underscore the functional roles of smORFs and relevance to human
60 diseases.

61

62 Advances in proteomics and next-generation sequencing (NGS) technologies have led
63 to a significantly improved annotation of smORF genes. There is evidence for >2,500
64 smORF genes in humans (Martinez et al., 2020), and there are over 1,000 annotated
65 smORF genes in the *Drosophila melanogaster* genome (Adams et al., 2000). smORF
66 genes are important regulators of physiology, development and metabolism and encode
67 hormones (Pearson et al., 1993), neurotransmitters (Snyder and Innis, 1979), ligands
68 and cofactors (Yang et al., 2015), RNA and DNA binding factors, and components of
69 ribonucleoproteins (Henras et al., 1998). Interestingly, a number of them are involved in

numerous mitochondrial functions and processes (Bosch et al., 2020; Rathore et al., 2018; Stein et al., 2018).

Although smORF genes are prevalent in metazoan genomes, a surprisingly small fraction of smORF genes are evolutionarily conserved in animals suggesting a high birth and death rate for these genes (Mackowiak et al., 2015; Martinez et al., 2020). For instance, there are over 2,500 smORF sequences with ribosome profiling evidence of translation across human cell lines, but only 273 of these human smORFs in the mouse genome based on computational analysis (Martinez et al., 2020). SmORF genes with deep evolutionary conservation are therefore of particular interest because of their assumed importance to the health and fitness across Metazoa and their implications for human health and disease.

Here, we provide an initial functional characterization of a collection of 194 conserved smORFs. For a subset of these smORFs, we describe their spatial expression patterns and phenotypes associated with gene loss of function or overexpression. Many of the pronounced phenotypes are associated with expression in neural tissue, and genes encoding mitochondrial proteins. In addition, several phenotypes were detectable only in flies subjected to stressful diets. This study serves as a resource for the functional annotation of this diverse and under-studied class of genes.

Results:

Deep conservation of smORFs

93 We identified 194 smORF genes that are evolutionarily conserved between humans and
94 *Drosophila melanogaster*, which we refer to as conserved smORFs (Materials and
95 Methods; Supplemental File 1). We further analyzed conservation of these 194
96 smORFs with the well-annotated transcriptomes of zebrafish (*Danio rerio*), nematodes
97 (*Caenorhabditis elegans)* and thale cress (*Arabidopsis thaliana)* (Figure 1A). There are
98 174 conserved in zebrafish, and 122 conserved in *C. elegans*. Notably, 68 conserved
99 smORFs were also conserved in *Arabidopsis*, providing evidence for the functional
100 importance of this dataset in non-Bilateria eukaryotes. Finally, amino acid alignment of
101 many of the human-fly conserved smORFs, such as *bc10*, *CG42497*, and *Tim10,* reveal
102 that they are also conserved among other invertebrate and vertebrate species (Figure
103 1B).
104
105 21 *Drosophila* conserved smORFs are polycistronic (Supplemental File 1, see example
106 in Figure 1B). Remarkably, while the individual smORFs that reside in *Drosophila*
107 polycistronic transcripts are evolutionarily conserved, their polycistronic structure is
108 generally not - indicating a complex evolutionary history. Interestingly, there are four
109 conserved smORFs encoded by polycistronic transcripts in both fly and *C. elegans*,
110 *CG42372, CG42375, CG44195,* and *Mocs2A*. Between flies and zebrafish, there are
111 two smORFs encoded by polycistronic transcripts in both species, *CG42497* and
112 *Mocs2A*. However, between flies and humans, or flies and *Arabidopsis,* there are none.
113 The human and *Arabidopsis* genomes are outliers in this analysis, wherein each
114 polycistronic *Drosophila* smORF is encoded within a single ORF transcript – consistent

115 with a model wherein transcriptional regulation is more complex in mammals and plants
116 (Lynch and Conery, 2003; Wray et al., 2003).
117
118 **Gene Ontology analysis of conserved smORFs**
119 The functions of conserved smORFs are diverse, as with any broad category of genes.
120 Gene Ontology (GO) Cellular Component enrichment analysis of conserved smORFs
121 determined that the majority of significantly enriched GO terms (Supplemental File 2)
122 are associated with mitochondrial function and localization (Figure 2, Supplemental File
123 3). Indeed, 58 conserved smORFs are predicted to be involved in mitochondrial function
124 (Supplemental File 3). "Mitochondrion" (P = 1.55 x 10^{-18}) which contains 36 conserved
125 smORFs is among the most significantly enriched GO Cellular Component terms in this
126 dataset, while "Mitochondrial Envelope" is the most significantly enriched GO Cellular
127 Component term (P = 1.07 x 10^{-21}). Additional significantly enriched terms include
128 mitochondrial inner membrane (P = 1.13 x 10^{-19}) and cytochrome complex (P = 2.80 x
129 10^{-11}) (Figure 2).
130
131 In contrast to the GO Cellular Component enrichment analysis, the most significantly
132 enriched Biological Function and Molecular Process GO terms (Supplemental File 2)
133 are related to serine-endopeptidase inhibitor activity (Figure 2; Supplemental File 3). For
134 instance, the most significantly enriched Molecular Function GO term is serine-type
135 endopeptidase inhibitor activity (P = 2.91 x 10^{-22}) (Figure 2). Notably, 30 of the 80
136 predicted serine-endopeptidase inhibitors in the *D. melanogaster* genome are in the
137 conserved smORF dataset. Of these 30 smORFs, 21 contain a predicted pancreatic

138 trypsin inhibitor Kunitz domain (Interpro: IPR002223), and nine smORFs contain a
139 predicted a Kazal domain (Interpro: IPR002350) - providing evidence that these 30
140 conserved smORFs are serine-endopeptidase inhibitors (Rawlings et al., 2004).
141 Additionally, all 30 of these predicted serine-endopeptidase inhibitors contain an N-
142 terminal secretion signal (SignalP6.0; 0.982 or higher score).
143
144 The oxidative phosphorylation pathway is the only significantly enriched pathway in the
145 smORF dataset (P = 3.58 x 10^{-14}; Kanehisa et al., 2021 (KEGG); Supplemental File 2,
146 Figure 2A, Figure S1). Interestingly, five of the 19 conserved smORFs that are assigned
147 to the oxidative phosphorylation pathway are not currently annotated as such (i.e., "CG"
148 IDs). Further analysis of these five conserved "CG" smORFs identified *CG9065* as the
149 fly *Cytochrome C oxidase subunit 17* (*COX17*). The remaining four, *CG31644*,
150 *CG40472, CG44296,* and *CG43206,* are duplicates of annotated fly genes: *cyclope*
151 (*cype*) encoding the *cytochrome c oxidase subunit 6C*, *NADH dehydrogenase*
152 *(ubiquinone) AGGG subunit* (*ND-AGGG*)*, Ubiquinol-cytochrome c reductase 6.4 kDa*
153 *subunit* (*UQCR-6.4*) and *Cytochrome c oxidase subunit 7C* (*COX7C*). Interestingly,
154 *CG31644, CG43206* and *CG44296* are primarily expressed in the adult testis, whereas
155 their paralogs (*cype*, *COX7C,* and *UQCR-6.4*, respectively*)* are ubiquitously expressed.
156
157 **In situ imaging of smORF mRNA expression in embryos**
158 To better understand the functional roles of smORFs in *Drosophila*, we performed in situ
159 mRNA hybridization to visualize smORF expression during embryogenesis
160 (Supplemental File 4; https://insitu.fruitfly.org/). Of these, organ-specific expression (i.e.,

161 “patterned” expression) could be assigned for 89 conserved smORFs during embryonic
162 development (Supplemental File 4).
163
164 In addition to patterned expression, conserved smORFs can also be classified as being
165 “maternally deposited”, and therefore ubiquitously expressed in the embryo in the
166 earliest developmental stage(s), and/or classified as being “ubiquitously expressed”,
167 where expression is observed throughout the full embryo after the earliest
168 developmental stage (Supplemental File 5). We found that 32 conserved smORFs were
169 classified as only being maternally deposited and/or being ubiquitously expressed (i.e.,
170 expressed throughout the entire embryo without assigned organ patterns) in at least
171 one embryonic stage. Of these 32 smORFs, *UQCR-6.4L* is the only smORF with no
172 observed expression after maternal deposition (i.e., after initial embryonic stages 1-3).
173
174 The remaining 72 smORFs revealed no observed expression in the embryo. In each
175 case this was concordant with RNA-seq (Brown, et al. 2018): these smORFs are
176 expressed later in development, and in specific tissues, or under stress conditions
177 (Brown, et al. 2018).
178
179 **In situ imaging of mitochondria-associated smORF mRNA expression**
180 Notably, mitochondria-associated conserved smORFs exhibited heterogenous spatial
181 expression patterns, indicating tissue- and stage-specific mitochondrial architectures,
182 and, by extension, tissue- and stage-specific mitochondrial functions (Figures 3A, B). Of
183 these, expression could be assigned as patterned for 34 smORFs. (Figures 3A, B). We

184 clustered these mitochondria-associated conserved smORFs with patterned expression
185 into five groups - showing extensive inter- and intra-group heterogeneity in organ
186 system expression patterns (Figures 3A, B).
187
188 Nearly all (33/34) patterned mitochondria-associated conserved smORFs were
189 maternally deposited (Supplemental File 4, Supplemental File 5), and 32/34 were
190 annotated as being ubiquitously expressed in at least one later embryonic stage
191 (Supplementary File 4). Of the two conserved smORFs without ubiquitous expression
192 after maternal deposition, one (*Cox17*) shows robust embryonic expression under RNA-
193 seq (Brown, et al. 2014.), but *in situ* hybridization was unsuccessful; while the other is
194 only annotated as patterned post maternal deposition (*CG17734*) (Supplemental File 4).
195 However, nearly all patterned, mitochondrial smORFs (33/34) exhibit complex and
196 tissue specific expression patterns after early embryonic ubiquitous expression (usually
197 due to maternal deposition) (Supplemental File 4). Hence, the zygotic regulation of
198 structural and functional components of mitochondria is highly patterned and specific to
199 individual tissues and cell types.
200
201 **Ribosome profiling and proteomics analysis to support the translation of**
202 **conserved smORFs**
203 To verify the translation of conserved smORFs, we performed ribosome profiling for six,
204 2-hour embryonic stages and six, 0-24 hr. mixed stage embryo samples (Figure 4;
205 Supplemental File 6, Supplemental File 7). We find evidence of translation for 91
206 conserved smORFs from these embryonic samples. We find that 22 (11%) conserved

207 smORFs are detected in only a single embryonic stage (Supplemental File 6),
208 consistent with transcriptional evidence from extensive RNA-seq experiments in a
209 developmental time course (Brown et al., 2014; Graveley et al., 2011).
210
211 We further analyzed published *Drosophila* proteomics datasets (Aradska et al., 2015;
212 Brown et al., 2018; Bucio-Mendez et al., 2019; Cammarato et al., 2011; Casas-Vila et
213 al., 2017; Dorus et al., 2006; Wasbrough et al., 2010) resulting in translation and/or
214 polypeptide support for 137 total conserved smORFs (Supplemental File 6). Analysis of
215 comprehensive *Drosophila* modENCODE mRNA expression data (Brown et al., 2014)
216 revealed 33/57 (58%) of conserved smORFs without evidence for translation or
217 peptides are maximally expressed in testes or accessory gland (Figure S2). Notably,
218 30/33 (91%) of these conserved smORFs that are maximally expressed in testes and
219 accessory gland are classified (FlyBase) as having no or low mRNA expression
220 throughout embryogenesis, and their expression is testes- and/or accessory gland-
221 specific.
222
223 **Functional analysis of smORF genes by F1 CRISPR screens**
224 Next, we assessed if conserved smORFs have important biological functions in
225 *Drosophila* by modifying their gene function in vivo. First, to knock out (KO) smORF
226 gene function in a systematic manner, we used a CRISPR/Cas9-based transgenic
227 crossing strategy (Port et al., 2020; Zirin et al., 2020) where a Cas9-expressing line is
228 crossed with sgRNA lines that target 5' coding sequence (sgRNA-KO). The resulting
229 progeny will contain somatic indels in the target gene that disrupt gene function. We

230 generated a collection of 177 sgRNA-KO lines that target 165 smORF genes
231 (Supplemental File 8). Each smORF sgRNA-KO line was crossed with *Act5c-Cas9*
232 (ubiquitous Cas9) and F1 progeny were screened for defects in viability, morphology, or
233 gross motor behavior (Figure 5A). Of the 115 sgRNA-KO lines tested, 14 (representing
234 14 genes) gave no mutant adult progeny or very few progenies compared to controls
235 (Figure 5B; Supplemental File 8), suggesting that they are essential genes. No other
236 morphological or behavioral phenotypes were observed for the remaining sgRNA-KO
237 lines.
238
239 To determine if the 14 putative essential smORFs play an important role in individual
240 tissues, we crossed the 14 sgRNA-KO hit lines to cell-type specific Cas9 lines (muscles,
241 gut enterocytes, dorsal thorax, wing disc, neurons) (Figure S3). Nearly all sgRNA-KOs
242 reduced viability when expressing Cas9 in neurons (12/14), whereas none reduced
243 viability with muscle Cas9 (0/14). Interestingly, a subset of sgRNA-KO lines had
244 reduced viability only when Cas9 was expressed in neurons (CG14057, CG40127,
245 CG14812, CG17776). In contrast, Rbp12 sgRNA-KO was lethal or showed low viability
246 with all Cas9 lines except muscle, and CG4650 sgRNA-KO did not cause any obvious
247 phenotypes with any of the six tissue-specific Cas9 lines. Finally, using a larval wing
248 disc Cas9 line, some sgRNA-KO lines caused adult wing defects, such as notching or
249 crumpled wings (Figure S3).
250
251 To over-express smORF genes in a systematic manner, we used CRISPR activation
252 (CRISPRa), where a sgRNA targets the promoter region and increases expression of

253 the endogenous gene (sgRNA-OE) via a catalytically dead Cas9 (dCas9) fused with a
254 transcriptional activator (e.g. VPR or SAM) (Ewen-Campen et al., 2017; Jia et al., 2018).
255 Similar to our sgRNA-KO collection, we generated a collection of 197 transgenic
256 sgRNA-OE lines that target 176 smORF genes. Each smORF sgRNA-OE line was
257 crossed with *tub-Gal4*, *UAS-dCas9-VPR* (abbreviated *tub>VPR*), and the F1 progeny
258 were screened for phenotypes as described for the sgRNA-KO collection (Figure 5C;
259 Supplemental File 8). Of the 123 sgRNA-OE lines tested, a small number had reduced
260 numbers of expected progeny, however none were statistically significant when
261 compared to control crosses (Figure 5D). Interestingly, CG13838 sgRNA-OE resulted in
262 viable adults that were flightless and had a “held-up” wing phenotype (Figure 5E). Both
263 phenotypes were 100% penetrant (n=100 flies). None of the remaining tested lines
264 produced aberrant morphological or behavioral phenotypes. To validate overexpression
265 by CRISPRa, we crossed 14 smORF-OE lines to *tub>VPR* and analyzed target gene
266 expression in adults using quantitative PCR (qPCR) (Figure S4). These results showed
267 that 6/14 sgRNA-OE lines had significantly elevated transcript expression when
268 normalized to *Rp49* or *Gapdh* and compared to negative control crosses (*attP40*).
269
270 To complement our results with CRISPRa, we overexpressed a subset of smORFs
271 using *UAS-cDNA* lines (Figure 5F; Supplemental File 8), which generally results in
272 higher levels of overexpression (Ewen-Campen et al., 2017). Of the 66 lines tested,
273 representing 61 genes, two lines were lethal, *UAS-CG18508* and *UAS-CG13838*. All
274 other lines had no obvious defects in viability, morphology, or behavior.
275

276 **Functional analysis of 25 uncharacterized smORF genes by whole animal KO**
277 F1 CRISPR screening tools are fast and scalable but have technical limitations, namely
278 that CRISPR-KO can produce mosaic phenotypes (Zirin et al., 2020) or phenotypes
279 outside the target tissue (Port and Bullock, 2016). Therefore, we wanted to apply a
280 more robust genetic tool to modify smORF gene function – whole animal knockout.
281 Since this requires greater time and resources, we targeted a subset of the conserved
282 smORFs with the least experimental characterization.
283
284 Using a combination of gene function prediction and manual searching (see Materials
285 and Methods), we identified 25 conserved smORF genes with minimal to no
286 experimental characterization in any organism with a corresponding homolog
287 (Supplemental File 8). Interestingly, 12 of these smORFs have a paralog (Figure S5A).
288 Using CRISPR/Cas9, we generated whole animal KOs for each of the 25 smORF genes
289 and multi-gene KOs for each paralog group (Supplemental File 8; Figure S5B-D). When
290 possible, we generated at least two independently derived alleles for each smORF. The
291 resulting homozygous animals were assessed for mutant phenotypes. Remarkably,
292 nearly all smORF KO lines were viable, fertile, and had no obvious morphological or
293 behavioral defects (Supplemental File 8). Interestingly, KO of the paralogs *CG32736*
294 and *CG42308* was lethal, either as single or double KO (Supplemental File 8; Figure
295 S5B-C). Characterization of *CG32736* and *CG42308* will be described elsewhere
296 (Bosch et al. 2020).
297

298 We reasoned that viable smORF KO mutants might reveal a mutant phenotype if raised
299 under stressful conditions. To test this, we transferred 24hr old homozygous smORF
300 KO embryos onto modified foods known to cause animal metabolic stress, starvation
301 (Zirin et al., 2015), high fat (Birse et al., 2010), and high salt (Stergiopoulos et al., 2009),
302 and measured developmental timing. On normal food, all tested KO mutants had similar
303 developmental timing compared to wild-type (Figure 6A). However, several KO mutants
304 exhibited significant developmental delays, low viability, or lethality on stressful foods
305 (Figure 6B-D). For example, two independent CG17931-KO alleles were lethal on
306 starvation food and were developmentally delayed or low viability on high salt food. In
307 addition, two independent alleles of CG42371-KO showed low viability on high salt food,
308 and one CG42371 allele had low viability on high fat food. Finally, bc10-KO was
309 developmentally delayed on high salt food.
310

311 **Discussion**

312 We identified 194 conserved smORFs between humans and fruit flies, the vast majority
313 of which are conserved broadly across bilaterians, with 68 conserved in *Arabidopsis*.
314 Ribosome profiling and prior proteomic experiments support the translation of 138 of
315 these smORFs. The remaining 59 without direct evidence of translation are largely
316 highly tissue specific (33 are specific to male reproductive tissues), and likely more
317 targeted samples are needed for detection. Of course, the thresholds we selected for
318 defining smORFs are somewhat arbitrary, and more lenient parameterizations (>100aa)
319 would yield more expansive lists. Certainly, this threshold produced an under-studied
320 collection of genes with diverse functions and phenotypes.

321

322 Of the 194 conserved smORFs, 21 are arranged in polycistronic transcripts. However,
323 this gene architecture is rarely conserved in distant species, with only four remaining
324 polycistronic in distant species – the remainder are partitioned into distinct
325 transcriptional units, often on different chromosomes. Polycistronic gene architectures
326 may have been selected against in eukaryotes – possibly as part of antiviral defense, as
327 mRNA viruses often make use of internal ribosome entry sites.

328

329 The largest class of conserved smORFs is related to mitochondrial structure and
330 function – including components of the oxidative phosphorylation pathway. Intriguingly,
331 mitochondrial smORFs exhibit highly tissue specific expression patterns after initial
332 ubiquitous maternal deposition. While mitochondrial functional diversity has been
333 explored in the nervous system (Fecher et al., 2019), this study indicates a far broader
334 diversity in mitochondrial architecture throughout the developing organism. The
335 profound conservation of these genes suggests that mitochondria are optimized in a
336 tissue-specific fashion throughout the bilaterian lineage. This observation points to an
337 evolutionary impetus for the translocation of mitochondrial genes to the host nuclear
338 genome – an intriguing direction for future study.

339

340 Results from our F1 CRISPR knockout screens revealed a number of essential smORF
341 genes. Interestingly, animal homologs of these gene hits (10/14) show lethality in other
342 organisms (Marrvel.org), and 5/14 are predicted mitochondrial (Supplemental File 3),
343 suggesting that fly lethality may be due to disrupted mitochondrial function.

344 Furthermore, these essential smORFs may be required in different tissues. For
345 example, Rbp12 knockout in either the gut, dorsal thorax, wing disc, or neurons caused
346 significant reduction in animal viability. In contrast, four essential smORFs were only
347 lethal when knocked out in neurons.
348
349 One interesting hit from our F1 knockout screen was *CG18508*. *CG18508* is the fly
350 homolog of *C18orf32*, which has been shown to associate with lipid droplets (Bersuker
351 et al., 2018). While ubiquitous *CG18508* knockout reduced viability, knockout in six
352 tissues did not. However, *CG18508* knockout in the wing disc caused adult wing
353 notching. Surprisingly, homozygous *CG18508* mutants are viable, fertile, and had
354 normal wing morphology, suggesting that *CG18508* sgRNA-KO has off-target effects.
355 Interestingly, overexpression of *CG18508* by UAS-cDNA was lethal. While it is not clear
356 if *CG18508* overexpression if physiologically relevant, it would be intriguing to examine
357 if animals die due to defects in lipid storage.
358
359 Results from the smORF overexpression screen revealed one additional smORF with
360 notable phenotypes. Overexpression of *CG13838* by CRISPRa resulted in flightless
361 adult flies with “held-up” wings, whereas overexpression by UAS-cDNA was lethal.
362 Since UAS-cDNA generally results in higher transcript expression compared to
363 CRISPRa (Ewen-Campen et al., 2017), *UAS-CG13838* cDNA lethality may be due to
364 higher expression. The *C. elegans* homolog of *CG13838*, *bubblin* (*bbln*), has recently
365 been characterized as essential for intermediate filament function (Remmelzwaal et al.,
366 2021). Interestingly, other fly mutations have been described that cause a wing

367 phenotype, such as *heldup*, which disrupts muscle thin filaments (Beall and Fyrberg,
368 1991). Like intermediate filaments, thin filaments are actin-based cytoskeletal structures
369 (Henderson et al., 2017). Therefore, *CG13838* overexpression may interfere with thin
370 filaments and/or intermediate filaments in muscle.
371
372 Many conserved smORFs have known or predicted functions. For example, smORF
373 CG14483 is uncharacterized in *Drosophila*, but its human homolog PET100 (65% aa
374 similarity, 37% aa identity) is a known regulator of mitochondrial complex IV biogenesis
375 and mutated in families with mitochondrial complex IV deficiency nuclear type 12
376 (MC4DN12) (Lim et al., 2014). In contrast, we identified 25 smORFs with little to no
377 characterization in any organism. For example, CG32736 is homologous to human
378 Small Integral Membrane Protein 4 (SMIM4) (66% aa similarity, 46% aa identity), but
379 had not been experimentally studied in any organism until recently (34969438, Bosch et
380 al. 2020). Studying poorly characterized smORF genes like *CG32736* could reveal new
381 biology and/or help understand human disease progression. Interestingly, we found
382 three previously uncharacterized smORFs (*CG17931/ SERFs*, *CG42371/ CEBPZOS*,
383 *bc10/BLCAP*) that were required for normal developmental progression on stressful
384 food diets.
385
386 **Acknowledgements:**
387 We thank Benjamin W Booth for processing the RNA-seq data and submitting RNA and
388 ribosome profiling data to the SRA. This work was funded by an award from the NIH
389 National Human Genome Research Institute (R01HG009352) to S.E.C. (Principal

390 Investigator). J.A.B. was supported by the Damon Runyon Foundation (DRG-2258-16)
391 and a “Training Grant in Genetics” T32 Ruth Kirschstein-National Research Service
392 Award institutional research training grant funded through the NIH/National Institute of
393 General Medical Sciences (T32GM007748). NP is an HHMI investigator. This article is
394 subject to HHMI’s Open Access to Publications policy. HHMI lab heads have previously
395 granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to
396 HHMI in their research articles. Pursuant to those licenses, the author-accepted
397 manuscript of this article can be made freely available under a CC BY 4.0 license
398 immediately upon publication.
399

400 **Materials and Methods:**

401 **Bioinformatic identification of 194 fly-human conserved smORFs**

402 First, 266,066 human smORFs were selected, including all annotated human smORF
403 transcripts (18,494 annotated in GENCODE version 24 less than 100aa), and 215,901
404 smORFs identified by 3-frame translations of all human transcripts that lack long ORFs
405 (11-100aa). This set was filtered to identify high-confidence elements by leveraging a
406 stringent, high-confidence set of conserved *Drosophila* smORFs. In *Drosophila*, there
407 are 984 unique peptides with no more than 100 aa from protein-coding genes annotated
408 at FlyBase (release 6.09) with evidence of translation. This set was expanded by taking
409 smORFs (11-100aa) predicted from two independent studies (Aspden et al. 2014),
410 adding 2,819 additional smORFs with evidence of translation from either Ribosome
411 Profiling or conservation among Drosophilidae. In addition, “stand alone” smORFs were

412 exclusively selected, and smORFs corresponding to either uORFs or smORFs that are
413 splice variants of longer ORFs were discarded.
414
415 The peptide sequences encoded by these smORFs in human and fly were compared
416 using BLASTP. Scrambled human peptide sequences were used as controls to
417 compute false discovery rates and determine how many good alignments one would
418 expect by chance alone. Conservation was further interrogated using Reverse PSI
419 BLAST (RPS-BLAST) to search for known, conserved protein domains in the human
420 and fly smORFs. By setting the E value cutoff on the BLASTP hits of 10^{-1} and 40%
421 smORF coverage (corresponding to an FDR ~ 5%), and further insisting that both the
422 human and the fly smORFs encode the same conserved domains (E-value Threshold
423 1e-5, FDR < 1%), 194 smORFs were identified that are conserved with human.
424 Because both conserved domains and BLASTP were used, the actual false discovery
425 rate is likely lower than 5%. To assess the completeness of our catalog based on
426 transcriptomes, a full 6-frame translation analysis of both genomes was conducted,
427 using BLASTP and RPS-BLAST to compare the millions of non-repetitive smORFs in
428 the human (10M) and fly (2.2M) genomes. Only seven new smORFs (new compared to
429 the 194 described above) passed our stringent conservation threshold, suggesting that
430 the actual FDR is much less than 5%, and that the 194 smORFs based on the
431 transcriptome is largely complete.
432
433 **Clustering of ontological anatomical annotations of embryonic expression**
434 **patterns**

435 Embryos were clustered using a bagged and cross validated procedure to ensure
436 cluster stability. The importance of stability was made apparent to us when, during the
437 course of our study, one additional smORF was discovered and added to this analysis,
438 and the resulting clusters differed significantly. We stabilized the clustering as follows:
439
440 First, we subsampled 80% of smORFs and used hierarchical clustering with Ward
441 linkage to form candidate clusters, holding out 20%. Here, the hold-out is to assess
442 stability by inducing some randomness -- note that this differs from the use of a holdout
443 in supervised learning. We selected 10K 80/20 splits. Cluster number was selected
444 using the Gap Statistic -- the first local maximum value was selected. We computed a
445 proximity matrix for consensus clustering as follows: for each pair of smORFs we
446 recorded the fraction of our 10K bagged clusterings in which they appeared in the same
447 cluster. This matrix was then input to hierarchical clustering with Ward linkage, and the
448 Gap Statistic was used to select cluster number as above.
449
450 **Embryo Collections for Coordinated Ribosome and RNA-Seq Profiling**
451 Embryos from ~14 g of Oregon-R flies were collected on standard molasses collection
452 trays after flies were acclimated to the environmental conditions of the cage (27℃ and
453 70% humidity) for three days. Six, two-hour embryonic time periods (0-2, 2–4 hr., 4–6
454 hr., 10-12 hr., 14–16 hr., and 16–18 hr. were collected simultaneously, which allowed
455 for immediate RNA-seq library and ribosome profiling construction of all six stages. See
456 Supplemental File 7 for detailed the ribosome profiling library protocol.
457 **RNA Preparation and Sequencing Methods**

458 Embryos were homogenized using a Pellet Pestle Cordless Motor (Kimble Cat. No.
459 749540-0000; Pellet Pestles Sigma Cat. No. Z359947), RNA was extracted using
460 TRIzol Reagent (Thermo Fisher, Cat. No. 15596026) and purified with the RNeasy Mini
461 Kit (QIAGEN Cat. No. 74106). Libraries were constructed with the NEBNext Ultra
462 Directional RNA Library Prep Kit for Illumina according to manufacturer's
463 recommendations (NEB, cat. no. E7420) using 14 cycles of PCR. Libraries were
464 sequenced on the Illumina NovaSeq 6000. Raw sequencing data is available at XXXX.
465 We used the STAR aligner v2.73a to align RNA-seq data to the *D. melanogaster*
466 genome (Rel 6). The picard 2.20.1 MarkDuplicates tool was used to remove PCR
467 duplicates and the deduplicated BAM alignment files were converted to bigWig format
468 using a custom tool and the UCSC bedGraphToBigWig tool. We ran fastp 0.20.1 to get
469 FASTQ file statistics.

470 **Ribosome Profiling Methods**

471 Polysome profiling was performed on all six samples. Briefly, embryos from each time
472 period (i.e., sample) were treated with harringtonine (LKT Laboratories, H0169)
473 in mild lysis buffer followed by the addition of cycloheximide and immediate grinding.
474 Samples were subjected to a 10%-50% sucrose gradient and all polysomes were
475 collected and combined. Collected polysome fractions were pelleted via sucrose
476 cushion (34%) and subjected to RNaseI digestion. After resuspension another cushion
477 (34%) was performed and RNA was coprecipitated with GlycoBlue (Thermo Fisher, cat.
478 No. AM95150). Recovered RNA was then run on a 15% TBE-Urea Gel (Thermo Fisher
479 cat. no. EC68852BOX), followed by gel size selection to isolate ribosome protected
480 fragments (26-31 nt) (Zymo Research, cat. no. R1070). Following end-repair

481 phosphorylation of RNA molecules, libraries were constructed with the NEBNext Small
482 RNA Library Preparation Kit according to manufacturer's recommendations (NEB, cat.
483 no. E7330). See Supplemental File 7 or detailed protocol.
484

485 **Sequencing, read processing and mapping**

486 Ribo-seq libraries were sequenced with the Illumina NovaSeq 6000. Raw sequencing
487 data is available at XXXX. Read processing and mapping were performed on an Ubuntu
488 18.04 Linux cluster running Kubernetes v1.16. on 240 total cores with 1500 GB of total
489 RAM. Reads were processed with the following commands:
490
491 I. QC of raw reads with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
492 fastqc file_one.fastq.gz file_two.fastq.gz
493
494 II. Clip adapters from Illumina raw reads with the following commands with Cutadapt
495 (Martin 2011)
496 cutadapt -a adapter sequence for file_one.fastq.gz \
497 -A adapter sequence for file_one.fastq.gz -j 10 \
498 -o file_one_clipped.fastq -p file_two_clipped.fastq \
499 file_one.fastq.gz file_two.fastq.gz
500
501 III. Trim clipped reads based on position quality with the following commands:
502 cutadapt -q 33 -j 12 -o file_one_clipped_trimmed.fastq file_one_clipped.fastq
503 cutadapt -q 33 -j 12 -o file_two_clipped_trimmed.fastq file_two_clipped.fastq
504
505 IV. QC of processed reads:
506 fastqc file_one_clipped_trimmed.fastq \ file_two_clipped_trimmed.fastq
507
508 V. Remove reads smaller than 26 bp and larger than 31 bp:
509 cutadapt --pair-filter=any --minimum-length=26 --maximum-length=31 -j 20 \
510 -o file_one_clipped_trimmed_min_max_removed.fastq \
511 -p file_two_clipped_trimmed_min_max_removed.fastq \
512 file_one_clipped_trimmed.fastq file_two_clipped_trimmed.fastq
513
514 VI. Map first to the rDNA reference with Bowtie2 (Langmead and Salzberg 2012) to 1)
515 remove rRNA sequences and decrease mapping time to nuclear reference genome, and
516 2) assess the percentage of rRNA contamination in each library:

```
Bowtie2 -x rDNA_reference_directory   \
-1 file_one_clipped_trimmed_min_max_removed.fastq   \
-2 file_two_clipped_trimmed_min_max_removed.fastq   \
--seedlen 12 --un-conc Bowtie2_mapping_directory -p 12  \
-S rDNA_mapping.sam
```

VII. Map non-rRNA reads to nuclear reference genome with STAR (Dobin et al. 2013):

```
STAR --runThreadN 25 --genomeDir --outFileNamePrefix  \
--outSAMtype BAM SortedByCoordinate   \
--winAnchorMultimapNmax 100 --seedSearchStartLmax 20  \
--outFilterMismatchNmax 3 --readFilesIn  \
Bowtie2_mapping_directory/un-conc-mate.1  \
Bowtie2_mapping_directory/un-conc-mate.2
```

Iterative Random Forest analysis of ORFquant Ribo-seq results to identify translated ORFs

Identification of translated sequences was performed using ORFquant (Calviello et al. 2020). For each predicted ORF with mapped reads we recorded all ORFquant summary statistics. This dataset was used as input for the iterative random forest (iRF) algorithm (Basu et al. 2018). A random forest model was trained with “embryonic non-translated genes” (lncRNAs and asRNAs) and “embryonic translated genes” (known protein coding genes expressed in embryonic stages). This trained iRF model was the applied to ORFquant output variables for each embryonic time period, producing a dataframe with translated or nontranslated calls for each *Drosophila* ORF.

Molecular biology

Fly genomic DNA was isolated by grinding a single fly in 50µl squishing buffer (10 mM Tris-Cl pH 8.2, 1 mM EDTA, 25 mM NaCl) with 200µg/ml Proteinase K (3115879001, Roche), incubating at 37˚C for 30 min, and 95˚C for 2 minutes. PCR was performed using Taq polymerase (TAKR001C, ClonTech) when running DNA fragments on a gel,

549 and Phusion polymerase (M-0530, NEB) was used when DNA fragments were
550 sequenced or used for molecular cloning. DNA fragments were run on a 1% agarose
551 gel for imaging or purified on QIAquick columns (28115, Qiagen) for sequencing
552 analysis. Sanger sequencing was performed at the DF/HCC DNA Resource Core facility
553 and chromatograms were analyzed using Lasergene 13 software (DNASTAR).
554
555 For isolating flies with frameshift indels in smORF genes, the target site was PCR
556 amplified from single fly genomic DNA and PCR fragments were Sanger sequenced.
557 For isolating flies with whole gene deletion of smORF genes by dual sgRNA cutting, the
558 target region was PCR amplified from single fly genomic DNA. Primers were designed
559 to flank the two sgRNA cut sites, such that a deletion of the intervening sequence would
560 produce a clear band size difference on an agarose gel. Deletion PCR fragments were
561 Sanger sequenced. Genotyping primer sequences are listed in Supplemental File 8.
562
563 For RT-qPCR analysis of smORF overexpression by CRISPRa, adult flies (tub-Gal4,
564 UAS-dCas9-VPR, sgRNA-OE) were flash frozen in liquid nitrogen. 4-14 frozen flies
565 (equal mixture of males and females) were homogenized in 1000ul Trizol (Invitrogen
566 15596026), RNA partially purified by chloroform extraction, and RNA extracted using a
567 Direct-zol RNA Miniprep kit (Zymo Research, R2050). cDNA was generated using the
568 iScript Reverse Transcription Supermix (BioRad 1708840). cDNA was analyzed by RT-
569 qPCR using iQ SYBR Green Supermix (BioRad 170-8880) on a CFX96 Real-Time
570 system (BioRad). qPCR primer sequences are listed in Supplemental File 8. Each
571 qPCR reaction was performed with five biological replicates (except *CG14818*, which

572 had two biological replicates), with two technical replicates each. Data from smORF
573 specific primers were normalized to primers that amplify GAPDH. Statistical significance
574 was calculated using a T-Test.
575
576 **Molecular cloning**
577 Plasmid DNAs were constructed and propagated using standard protocols. Briefly,
578 chemically competent TOP10 E.coli. (Invitrogen, C404010) were transformed with
579 plasmids containing either Ampicillin or Kanamycin resistance genes and were selected
580 on LB-Agar plates with 100µg/ml Ampicillin or 50µg/ml Kanamycin. Oligo sequences are
581 in Supplemental File 8.
582
583 sgRNA expression plasmids: Plasmids encoding sgRNAs were generated using
584 previously described protocols. For sgRNAs cloned into *pCFD3* (Port et al. 2014),
585 annealed oligos encoding a sgRNA spacer were ligated into *pCFD3* digested with BbsI
586 (NEB, R3539). For sgRNAs cloned into *pCFD4* (Port et al. 2014), dual sgRNAs were
587 PCR amplified from *pCFD4* template, and inserted by Gibson assembly (NEB, E2611)
588 into *pCFD4* digested with BbsI. For sgRNAs cloned into *pCFD5* (Port and Bullock
589 2016), dual sgRNAs were PCR amplified from *pCFD5* template, and inserted by Gibson
590 assembly into *pCFD5* digested with BbsI. sgRNAs cloned into *pCFD3* and were
591 performed by DRSC/TRiP (https://fgr.hms.harvard.edu/). Information on sgRNA-KO and
592 sgRNA-OE plasmids generated by the DRSC/TRiP is available at
593 https://www.flyrnai.org/tools/grna_tracker/web/. Information on remaining *pCFD4* and
594 *pCFD5* sgRNA plasmids is in Supplemental File 8.

595

596 pEntr plasmids: Entry plasmids were generated by PCR amplifying coding sequence
597 and inserting into *pEntr* using either dTopo (Invitrogen, K240020) or Gibson assembly
598 (NEB, E2611). cDNA sequence was PCR amplified from cDNA reverse transcribed from
599 total RNA (either S2R+ cell or adult fly) or adult fly genomic DNA.

600

601 Gateway cloning LR reactions: Gateway cloning reactions were performed using LR
602 Clonase II Enzyme mix (Invitrogen 11791-020). pEntr_cDNA plasmids were recombined
603 with *pWalium10-roe* (Perkins et al. 2015).

604

605 **Fly Genetics**

606 Flies were maintained on standard fly food at 25˚C unless otherwise noted.

607

608 Fly stocks were obtained from the Perrimon lab collection, Bloomington Stock center
609 (indicated with BL#), or generated in this study (see below).

610

611 Perrimon Lab stocks:

612 *yv; Gla/CyO*

613 *yw; nos-Cas9attP40/CyO*

614 *yw;; nos-Cas9attP2*

615 *lethal/FM7,GFP*

616 *yw; Gla/CyO*

617 *yw;; TM3, Sb/TM6b*

618 *lethal/FM7,GFP;; TM3, Ser*

619 *yw; Sp/CyO; MKRS/TM6B*

620

621 Bloomington Stocks:

622 sgRNA lines (see Supplemental File 8)

623 *UAS-dCas9-VPR; tub-Gal4/S-T* (*tub>VPR*) (BL67048)

624 *Actin-Cas9* (BL54590)

625 *attP40* (BL36304)

626 *Mhc>Cas9* (BL67079)

627 *LSP>Cas9* (BL67087)

628 *Myo1a>Cas9* (BL67088)

629 *Pnr>Cas9* (BL67077)

630 *Nub>Cas9* (BL67086)

631 *Elav>Cas9* (BL67073)

632 *yw* (BL1495)

633

634 Information on the smORF KO alleles are described in Supplemental File 8.

635

636 <u>smORF knockout by frameshift indel by transgenic crossing:</u> Flies expressing a sgRNA

637 that targets the 5' coding sequence (see Supplemental File 8) were crossed with *nos-*

638 *Cas9* flies. *nos-Cas9attP2* was used for targeting genes on Chromosomes X and II, and

639 *nos-Cas9attP40* was used for targeting genes on chromosome III. F1 progeny were

640 crossed with a balancer strain. Single fly F2 progeny were crossed with a balancer

641 strain, taken for genotyping, and F2 crosses with a frameshift indel were kept and
642 balanced and homozygosed if possible. Frameshift knockout lines were generated
643 either by WellGenetics, Shu Kondo, or in the Perrimon lab.
644
645 smORF knockout by full gene deletion by injection: Plasmids encoding two sgRNAs that
646 flank a gene locus were injected into *nos-Cas9* embryos. Injected F0 adults were
647 crossed to with a balancer strain. Single fly F1 progeny were crossed with a balancer
648 strain, taken for genotyping, and F1 crosses with a full gene deletion were kept and
649 balanced and homozygosed if possible.
650
651 smORF knockout by CRISPaint insertion by injection: To generate the bc10 KO allele, a
652 plasmid encoding a sgRNA that targets the 5' coding sequence of bc10 (GP01409) was
653 co-injected with pCRISPaint-T2A-Gal4-3xP3-RFP (Addgene #127556) into *nos-*
654 *Cas9attP40* embryos. Injected F0 adults were outcrossed to *yw*, single RFP+ F1
655 progeny were crossed with *yw;; TM3, Sb/TM6b,* and the insertion in bc10 was verified
656 by PCR and sanger sequencing. The *bc10-CRISPaint* allele contains *T2A-Gal4* inserted
657 in the reverse orientation relative to the 5'-3' *bc10* transcript, and thus is not a Gal4
658 reporter allele.
659
660 To generate double smORF knockout lines, smORF alleles on different chromosomes
661 were brought into the same strain by outcrossing to double balancer lines.
662

663 CRISPR-KO F1 crosses and phenotyping: Lines expressing sgRNAs that target smORF
664 5' coding sequence were crossed with a line ubiquitously expressing Cas9, *Act5c-Cas9*.
665 Specifically, male sgRNA flies were used that were heterozygous with a balancer
666 chromosome (CyO or TM3, Sb), and were crossed with homozygous *Act5c-Cas9*
667 female flies. To quantify the viability of F1 flies with somatic KO, we recorded the
668 number of balancer progeny (*Act5c-Cas9/Bal*) and non-balancer progeny
669 (*sgRNA/Act5c-Cas9*). The total number of F1 progeny counted per cross was
670 920>n>31. To calculate a viability score, the number non-balancer flies was divided by
671 the mendelian expected number of non-balancer flies (# F1 progeny/2) and multiplied
672 by 100 (# observed non-balancer/# expected non-balancer*100). Negative control
673 crosses were *attp40/CyO* males crossed with *Act5c-Cas9* females. For crosses leading
674 to reduced viability, a Chi-square test was used to determine significance by comparing
675 the # expected non-balancer flies from negative control (attp40) vs experimental
676 (sgRNA) crosses. Those sgRNA hits that had significant reduced viability (Figure 5,
677 Supplemental File 8) were crossed with tissue specific Cas9 lines (Figure S3). Viability
678 scores and Chi-square tests were performed similarly to *Act5c-Cas9* crosses.
679
680 CRISPRa F1 crosses and phenotyping: Lines expressing sgRNAs that target upstream
681 of a smORF transcriptional start site (TSS) were crossed with line ubiquitously
682 expressing dCas9-VPR, *tub>VPR* (*tub-Gal4, UAS-dCas9-VPR/S-T*). "S-T" are a second
683 and third chromosome balancer pair that segregate together due to a reciprocal
684 translocation, and are marked by Cy, Hu, and Tb (T(2;3)TSTL14, SM5: TM6B, Tb[1]).
685 Male sgRNA flies were crossed with tub>VPR females. A viability score was calculated

686 similar to CRISPR-KO F1 crosses, (# observed non-balancer/# expected non-
687 balancer*100). When using homozygous sgRNA males, the expected number of non-
688 balancer flies was # F1 progeny/2. When using heterozygous sgRNA/Bal males, the
689 expected number of non-balancer flies was # F1 progeny/4. The total number of F1
690 progeny counted per cross was 346>n>56. Negative control crosses were *attp40/CyO*
691 males crossed with *tub>VPR* females. Chi-square analysis was performed similarly to
692 CRISPR-KO F1 crosses. No crosses had significantly reduced viability.
693
694 <u>cDNA overexpression crosses and phenotyping:</u> Transgenic UAS-cDNA lines were
695 crossed with a ubiquitous Gal4 line. For convenience, we used the same driver used for
696 CRISPRa crosses, *tub>VPR* (*tub-Gal4, UAS-dCas9-VPR/S-T*). A viability score was
697 calculated similar to CRISPRa F1 crosses. Negative control crosses were *attp40/CyO*
698 males crossed with *tub>VPR* females. The total number of F1 progeny counted per
699 cross was 706>n>100. Chi-square analysis was performed similarly to CRISPR-KO F1
700 crosses.
701
702 <u>Stressful food recipes:</u> For all food types, standard lab fly food was melted in a
703 microwave, and distilled water (dH20) was added as 10% boiled volume (100ml boiled
704 food + 10ml dH20) to replace the evaporated water. This is used as control food. For
705 high salt food, solid NaCl (Fisher Scientific, S271) was added at 30% weight per volume
706 of control food (e.g. 100ml control food + 30g NaCl) and mixed well. For high fat food,
707 solid coconut oil (Sigma, W530155) was added at 30% weight per volume of control
708 food (e.g. 100ml control food + 30g coconut oil) and mixed well. For starvation food,

709 control food was diluted to 30% in 1% melted agar (BD, 214030) in 1x PBS (Gibco,
710 10010-023) (e.g. 30ml control food + 70ml melted 1% agar in 1x PBS). Melted liquid
711 food types were poured into empty vials and cooled at 4˚C.
712
713 Quantification of developmental timing on stressful food: 24hr larvae from homozygous
714 viable smORF lines were transferred to stressful food or control food and raised on this
715 food until pupal eclosion as adults. To increase the fecundity of the adult flies, four days
716 prior to the 24hr larval collection, ~150 adult flies from each KO line were transferred to
717 fresh food containing yeast paste. Adult flies were transferred onto fresh food bottles
718 containing yeast paste and allowed to lay for 4hr. 24hr after the end of egg deposition,
719 30 freshly hatched larvae (24hr-28hr old) were transferred into vials containing either
720 control food or stressful food (control, high fat, high salt, starvation). The time of
721 pupariation and fly eclosion was determined once at least 15 flies pupariated and
722 eclosed, respectively. yw flies were included as a negative control genotype. Each
723 genotype-foodtype experiment was carried out in at least triplicate. For those genotype-
724 foodtypes with a developmental delay, significance was calculated using a One-Way
725 ANOVA test run with a Dunnet post-hoc test using GraphPad Prism.
726
727 Adult wing mounting, imaging, analysis: sgRNA-KO lines were crossed with *nub>Cas9*
728 and the wings of adult progeny were removed using forceps under a dissecting
729 microscope. For each genotype, at least 6 wings were collected. Removed wings were
730 placed onto a drop of mounting medium (50% Permount (Fisher Scientific, SP15), 50%
731 Xylenes (Fisher Scientific X5)) on a microscope slide (Thermo Scientific, 3050) and

732 mounted using a coverslip (VWR, 48393059). The coverslip was sealed to the
733 microscope slide with clear nail polish. Images of the wings were taken using a stereo
734 microscope (Zeiss Axio Zoom V16) at 32x magnification. The area of each wing was
735 measured using the polygon function in ImageJ. The area measurements were
736 analyzed in Graphpad Prism. A One-way ANOVA test was carried out with a Dunnett
737 Post-Hoc test to determine the significance of the area of the wings of the experimental
738 cross progeny when compared with the control cross progeny (*attP40 X nub>Cas9)*.
739

740 **Bioinformatics and literature searching**

741 For protein alignments in Figure S5, we downloaded protein sequence files from
742 Flybase.org or NCBI, aligned them using Clustal Omega
743 (https://www.ebi.ac.uk/Tools/msa/clustalo), and generated alignment images using
744 JalView (https://www.jalview.org/).
745

746 For literature searching for smORF homolog characterization, we queried every
747 ultraconserved *Drosophila* smORF using the following online tools: Gene2function
748 (Gene2function.org), HUGO Gene Nomenclature Committee (genenames.org), Interpro
749 (ebi.ac.uk/interpro), and Alliance of Genome Resources (alliancegenome.org). Those
750 smORF genes (25) that had no or minimal characterization in orthologs were selected.
751

752 **Figure Legends**

753 **Figure 1. Conservation of smORF dataset. (*A*)** Number of conserved smORFs in
754 dataset with and without homologs in a selection of species with well-annotated

755 transcriptomes. ***(B)*** Multiple-species alignments of conserved smORFs, including mean
756 amino acid hydrophobicity at each alignment position. *bc10* (upper transcript) is
757 encodes one smORF, whereas *CG42497* and *Tim10* (lower transcript) is polycistronic.
758
759 **Figure 2. Gene Ontology (GO) and KEGG enrichment analysis of conserved**
760 **smORFs.** Significantly enriched GO terms for molecular function ("MF"), biological
761 process ("BP"), cellular component ("CC") are plotted. GO and KEGG enrichment
762 analyses were performed with g:Profiler (Raudvere et al. 2019). Significantly enriched
763 terms $<10^{-5}$ are shown that also encompass all conserved smORFs classified as serine-
764 type endopeptidase inhibitors and mitochondria-associated conserved smORFs.
765
766 **Figure 3. *In situ* mRNA hybridization patterns of mitochondria-associated**
767 **smORFs. (*A*)** Clustering of mitochondria-associated conserved smORF *in situ* mRNA
768 expression patterns (Supplemental Methods). For each mitochondrial conserved
769 smORF, the organs where expression patterns were assigned are represented by red
770 boxes, while blue boxes represent no annotated expression. Expression patterns across
771 embryo stages are collapsed. "DPPM" = dorsal prothoracic pharyngeal muscle; "MT" =
772 Malpighian tubules; "VNC" = ventral nerve cord; "D and V epidermis" = dorsal and
773 ventral epidermis. **(*B*)** In-situ mRNA hybridization images for each mitochondria-
774 associated conserved smORF with patterned expression. Each image was taken
775 between embryonic stages 13-16.
776

777 **Figure 4. Ribosome Profiling.** ***(A)*** Overview of ribosome profiling workflow, where
778 polysomes are isolated followed by digestion of inter-ribosome RNA. Ribosome
779 protected fragments are then collected. **(*B*)** Recovered RNA was run on a page urea gel
780 and RNA molecules in the range of ribosome protected fragments (RPFs) are excised
781 and purified. After sequencing, the number of in frame reads are analyzed **(*C*)** to
782 determine if RPFs were successfully sequenced. **(*D*)** Distribution of tags per million
783 (TPM) for six, embryonic time periods. Comparison of ribosome profiling sequencing to
784 mRNA-seq sequencing showing ribosome profiling libraries are constrained to CDS
785 while mRNA libraries map to the entire annotated transcript.
786
787 **Figure 5. Functional characterization of conserved smORFs by F1 CRISPR in vivo**
788 **screening. (*A*)** Genetic cross to perform CRISPR somatic knockout in F1 generation.
789 **(*B*)** Quantification of viability of F1 flies from 115 sgRNA-KO crosses. Number of F1
790 progeny counted per cross was 918>n>33. **(*C*)** Genetic cross to perform CRISPR gene
791 overexpression in F1 generation. ***(D)*** Quantification of viability of F1 flies from 123
792 sgRNA-OE crosses. Number of F1 progeny counted per cross was 220>n>56. **(E)**
793 Images of adult female flies aged seven days after eclosion for two indicated genotypes.
794 **(F)** Quantification of viability of F1 flies from 68 UAS-cDNA crosses. Number of F1
795 progeny countered per cross was 706>n>101.
796
797 **Figure 6. Full CRISPR knockout of 25 uncharacterized smORFs raised on**
798 **stressful foods.** Developmental timing of egg deposition to adult eclosure in smORF
799 KO mutants raised on ***(A)*** control food, ***(B)*** high salt food (30% NaCl), ***(C)*** high fat food

800 (30% coconut oil), ***(D)*** starvation food (30% food, 70% PBS+1%agar). Significance was

801 determined by One Way Anova test followed by a Dunnet Post-Hoc, P=0.05(*),

802 P=0.01(**), P=0.001(***), error bars calculated with SD.

803

804

805 **References**

806 Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G.,
807 Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F.*, et al.* (2000). The genome sequence
808 of *Drosophila melanogaster*. *Science 287*, 2185-2195.
809
810 Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MA, Brocard M, Couso JP.
811 Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife.
812 2014 Aug 21;3:e03528. doi: 10.7554/eLife.03528. PMID: 25144939; PMCID:
813 PMC4359375.
814
815 Aradska, J., Bulat, T., Sialana, F.J., Birner-Gruenberger, R., Erich, B., and Lubec, G.
816 (2015). Gel-free mass spectrometry analysis of *Drosophila melanogaster* heads.
817 *Proteomics 15*, 3356-3360.
818
819 Basrai, M.A., Hieter, P., and Boeke, J.D. (1997). Small open reading frames: beautiful
820 needles in the haystack. *Genome Res 7*, 768-771.
821
822 Basu S, Kumbier K, Brown JB, Yu B. 2018. Iterative random forests to discover
823 predictive and stable high-order interactions. *Proc Natl Acad Sci U S A* 115:1943-1948.
824
825 Beall, C.J., and Fyrberg, E. (1991). Muscle abnormalities in *Drosophila melanogaster*
826 heldup mutants are caused by missing or aberrant tropinin-I isoforms. *J Cell Biol 114*,
827 941-951.
828
829 Bersuker, K., Peterson, C.W.H., To, M., Sah, S.J., Savikhin, V., Grossman, E.A.,
830 Nomura, D.K., Olzmann, J.A. (2018). A proximity labeling strategy provides insights into
831 the composition and dynamics of lipid droplet proteomes. *Dev Cell 44*, 97-112.
832
833 Birse, R.T., Choi, J., Reardon, K., Rodriguez, J., Graham, S., Diop, S., Ocorr, K.,
834 Bodmer, R., and Oldham, S. (2010). High-fat-diet-induced obesity and heart dysfunction
835 are regulated by the TOR pathway in *Drosophila*. *Cell Metab 12*, 533-544.
836
837 Bosch, J., Ugur, B., Pichardo-Casas, I., Rabasco, J., Escobedo, F., Zuo, Z., Brown,
838 J.B., Celniker, S., Sinclair, D., Bellen, H.*, et al.* (2020). Two neuronal peptides encoded
839 from a single transcript regulate mitochondrial function in *Drosophila*. *BioRxiv*.

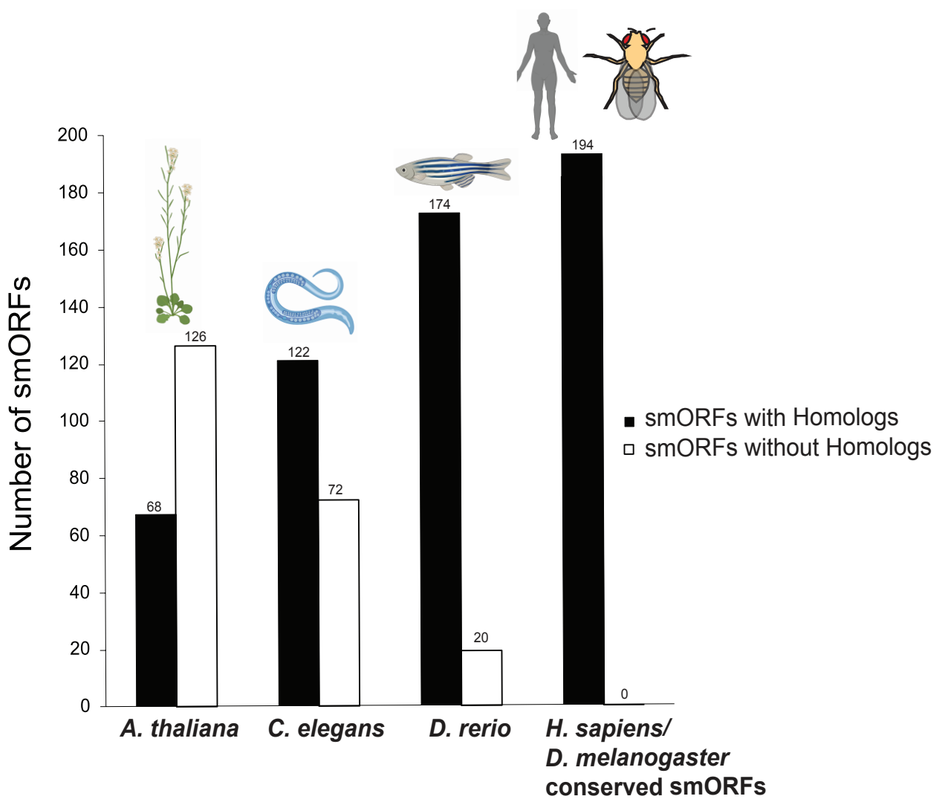
840
841 Brown, C.J., Kaufman, T., Trinidad, J.C., and Clemmer, D.E. (2018). Proteome changes
842 in the aging *Drosophila melanogaster* head. *Int J Mass Spectrom 425*, 36-46.
843
844 Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W.,
845 Wen, J., Park, S., Suzuki, A.M., et al. (2014). Diversity and dynamics of the *Drosophila*
846 transcriptome. *Nature* 512, 393-399.
847
848 Bucio-Mendez, A., Cruz-Becerra, G., Valadez-Graham, V., Dinkova, T.D., and Zurita, M.
849 (2019). The Dmp8-Dmp18 bicistron messenger RNA enables unusual translation during
850 cellular stress. *J Cell Biochem 120*, 3887-3897.
851
852 Cammarato, A., Ahrens, C.H., Alayari, N.N., Qeli, E., Rucker, J., Reedy, M.C., Zmasek,
853 C.M., Gucek, M., Cole, R.N., Van Eyk, J.E.*, et al.* (2011). A mighty small heart: the
854 cardiac proteome of adult *Drosophila melanogaster*. *PLoS One 6*, e18497.
855
856 Casas-Vila, N., Bluhm, A., Sayols, S., Dinges, N., Dejung, M., Altenhein, T., Kappei, D.,
857 Altenhein, B., Roignant, J.Y., and Butter, F. (2017). The developmental proteome of
858 *Drosophila melanogaster*. *Genome Res 27*, 1273-1285.
859
860 Calviello L, Hirsekorn A, Ohler U. 2020. Quantification of translation uncovers the
861 functions of the alternative transcriptome. *Nat Struct Mol Biol* 27:717-725.
862
863 Couso, J., Patriquim, P. (2017). Classification and function of small open reading
864 frames. *Nat Rev Mol Cell Biol 9*, 575-589.
865
866 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. 2013. STAR:
867 Ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.
868
869 Dorus, S., Busby, S.A., Gerike, U., Shabanowitz, J., Hunt, D.F., and Karr, T.L. (2006).
870 Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nat*
871 *Genet 38*, 1440-1445.
872
873 Ewen-Campen, B., Yang-Zhou, D., Fernandes, V.R., Gonzalez, D.P., Liu, L.P., Tao, R.,
874 Ren, X., Sun, J., Hu, Y., Zirin, J.*, et al.* (2017). Optimized strategy for in vivo Cas9-
875 activation in *Drosophila*. *Proc Natl Acad Sci U S A 114*, 9409-9414.
876
877 Fecher, C., Trovo, L., Muller, S.A., Snaidero, N., Wettmarshausen, J., Heink, S., Ortiz,
878 O., Wagner, I., Kuhn, R., Hartmann, J.*, et al.* (2019). Cell-type-specific profiling of brain
879 mitochondria reveals functional and molecular diversity. *Nat Neurosci 22*, 1731-1742.
880
881 Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri,
882 C.G., van Baren, M.J., Boley, N., Booth, B.W.*, et al.* (2011). The developmental
883 transcriptome of *Drosophila melanogaster*. *Nature 471*, 473-479.
884

885 Hammonds, A.S., Bristow, C.A., Fisher, W.W., Weiszmann, R., Wu, S., Hartenstein, V.,
886 Kellis, M., Yu, B., Frise, E., and Celniker, S.E. (2013). Spatial expression of transcription
887 factors in *Drosophila* embryonic organ development. *Genome Biol 14*, R140.
888
889 Henderson, C.A., Gomez, C.G., Novak, S.M., Mi-Mi, L., Gregorio, C.C. (2017).
890 Overview of the muscle cytoskeleton. *Compr Physiol 7,* 891-944.
891
892 Henras, A., Henry, Y., Bousquet-Antonelli, C., Noaillac-Depeyre, J., Gelugne, J.P., and
893 Caizergues-Ferrer, M. (1998). Nhp2p and Nop10p are essential for the function of
894 H/ACA snoRNPs. *EMBO J 17*, 7078-7090.
895
896 Jia, Y., Xu, R.G., Ren, X., Ewen-Campen, B., Rajakumar, R., Zirin, J., Yang-Zhou, D.,
897 Zhu, R., Wang, F., Mao, D.*, et al.* (2018). Next-generation CRISPR/Cas9 transcriptional
898 activation in *Drosophila* using flySAM. *Proc Natl Acad Sci U S A 115*, 4719-4724.
899
900 Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2021).
901 KEGG: integrating viruses and cellular organisms*. Nucleic Acids Res 49*, D545-D551.
902
903 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nat*
904 *Methods* 9:357-359.
905
906 Larkin A, Marygold SJ, Antonazzo G, Attrill H, Dos Santos G, Garapati PV, et al. 2021.
907 FlyBase: Updates to the *Drosophila melanogaster* knowledge base. *Nuc Acids Res*
908 49:D899-D907.
909
910 Lim, S.C., Smith, K.R., Stroud, D.A., Compton, A.G., Tucker, E.J., Dasvarma, A.,
911 Gandolfo, L.C., Marum, J.E., McKenzie, M., Peters, H.L.*, et al.* (2014). A founder
912 mutation in PET100 causes isolated complex IV deficiency in Lebanese individuals with
913 Leigh syndrome. *Am J Hum Genet 94*, 209-222.
914
915 Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. *Science 302*,
916 1401-1404.
917
918 Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni,
919 G., Rajewsky, N., Kempa, S., Selbach, M.*, et al.* (2015). Extensive identification and
920 analysis of conserved small ORFs in animals. *Genome Biol 16*, 179.
921
922 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput
923 sequencing reads. *EMBnet* 17:10-12.
924
925 Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A.
926 (2020). Accurate annotation of human protein-coding small open reading frames. *Nat*
927 *Chem Biol 16*, 458-468.
928

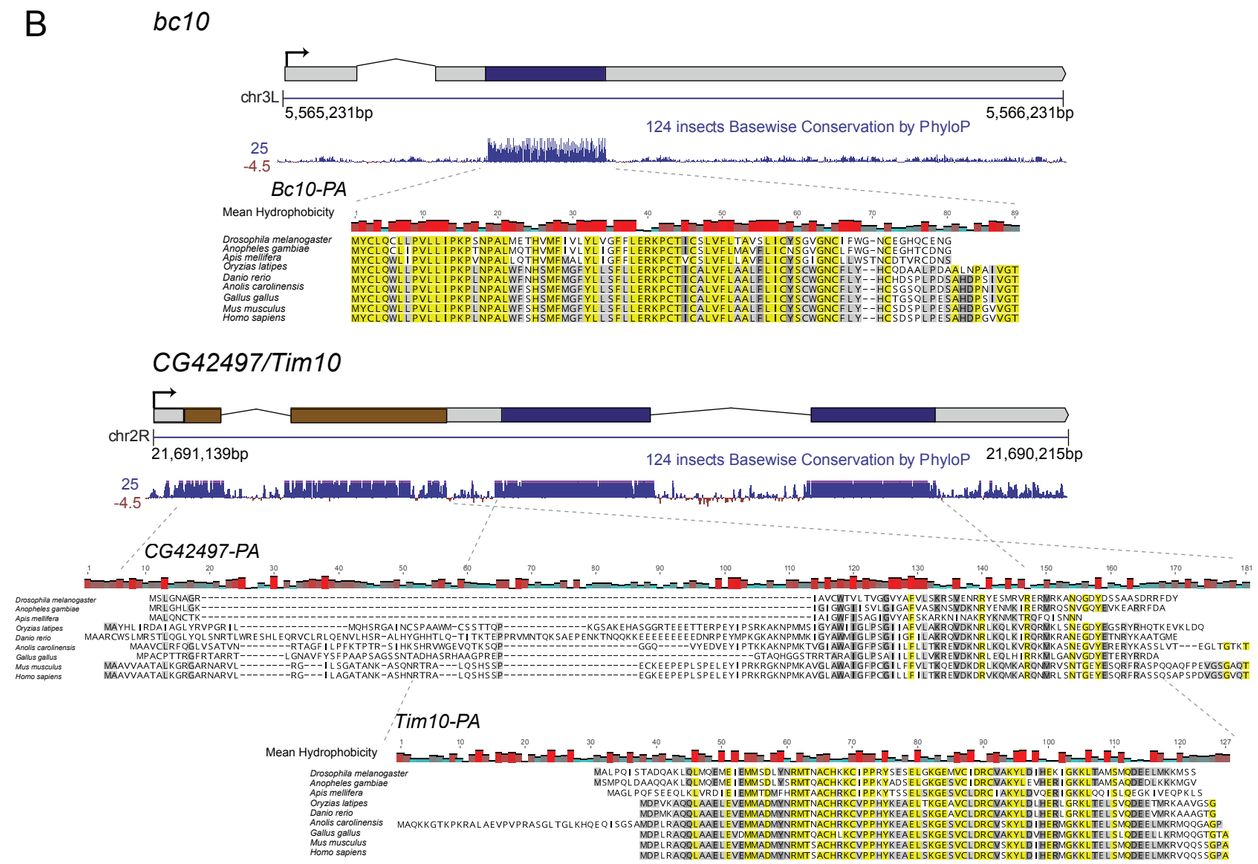
929 Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018). An update on sORFs.org:
930 a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res 46*, D497-
931 D502.
932
933 Pearson, R.K., Anderson, B., and Dixon, J.E. (1993). Molecular biology of the peptide
934 hormone families. *Endocrinol Metab Clin North Am 22*, 753-774.
935
936 Perkins LA, Holderbaum L, Tao R, Hu Y, Sopko R, McCall K, Yang-Zhou D, Flockhart I,
937 Binari R, Shim HS, Miller A, Housden A, Foos M, Randkelv S, Kelley C, Namgyal P,
938 Villalta C, Liu LP, Jiang X, Huan-Huan Q, Wang X, Fujiyama A, Toyoda A, Ayers K,
939 Blum A, Czech B, Neumuller R, Yan D, Cavallaro A, Hibbard K, Hall D, Cooley L,
940 Hannon GJ, Lehmann R, Parks A, Mohr SE, Ueda R, Kondo S, Ni JQ, Perrimon N. The
941 Transgenic RNAi Project at Harvard Medical School: Resources and Validation.
942 Genetics. 2015 Nov;201(3):843-52. doi: 10.1534/genetics.115.180208. Epub 2015 Aug
943 28. PMID: 26320097; PMCID: PMC4649654.
944
945 Plaza, S., Menschaert, G., Payre, F. (2017). In search of lost small peptides. *Annu Rev*
946 *Cell Dev Biol 33* 391-416.
947
948 Port F, Chen HM, Lee T, Bullock SL. Optimized CRISPR/Cas tools for efficient germline
949 and somatic genome engineering in Drosophila. Proc Natl Acad Sci U S A. 2014 Jul
950 22;111(29):E2967-76. doi: 10.1073/pnas.1405500111. Epub 2014 Jul 7. PMID:
951 25002478; PMCID: PMC4115528.
952
953 Port, F., and Bullock, S.L. (2016). Augmenting CRISPR applications in *Drosophila* with
954 tRNA-flanked sgRNAs. *Nat Methods 13*, 852-854.
955
956 Port, F., Strein, C., Stricker, M., Rauscher, B., Heigwer, F., Zhou, J., Beyersdorffer, C.,
957 Frei, J., Hess, A., Kern, K.*, et al.* (2020). A large-scale resource for tissue-specific
958 CRISPR mutagenesis in *Drosophila*. *eLife* 9, e53865.
959
960 Rathore, A., Chu, Q., Tan, D., Martinez, T.F., Donaldson, C.J., Diedrich, J.K., Yates,
961 J.R., 3rd, and Saghatelian, A. (2018). MIEF1 microprotein regulates mitochondrial
962 translation. *Biochemistry 57*, 5564-5575.
963
964 Rawlings, N.D., Tolle, D.P., Barrett, A.J. (2004). Evolutionary families of peptidase
965 inhibitors. *Biochem J 378,* 705-716.
966
967 Remmelzwaal, S., Geisler, F., Stucchi, R., van der Horst, Suzanne., Pasolli, M., Kroll,
968 J.R., Jarosinska, O.D., Akhmanova, A., Richardson, C.A., Altelaar, M., Leube, R.E.,
969 Ramalho, J.J., Boxem, M. (2021). BBLN-1 is essential for intermediate filament
970 organization and apical membrane morphology. *Curr Biol 31*, 2334-2346.
971
972 Snyder, S.H., and Innis, R.B. (1979). Peptide neurotransmitters. *Annu Rev Biochem 48*,
973 755-782.
974

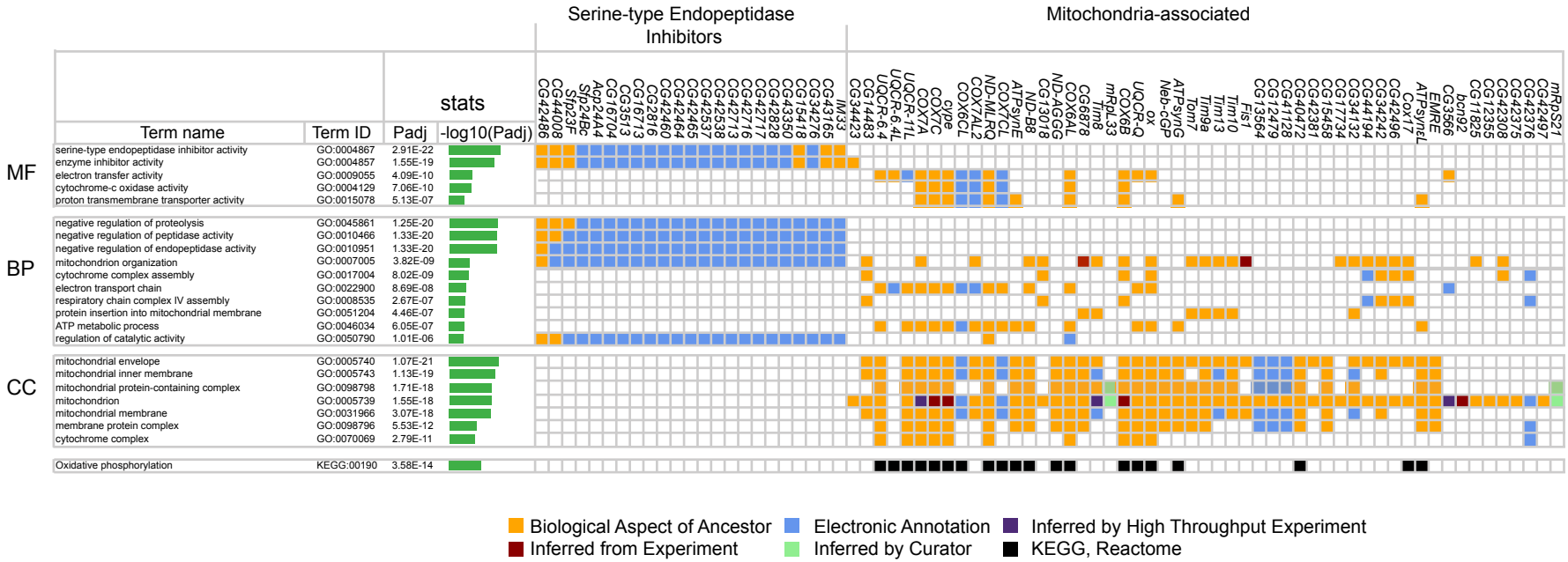
975 Stein, C.S., Jadiya, P., Zhang, X., McLendon, J.M., Abouassaly, G.M., Witmer, N.H.,
976 Anderson, E.J., Elrod, J.W., and Boudreau, R.L. (2018). Mitoregulin: a lncRNA-encoded
977 microprotein that supports mitochondrial supercomplexes and respiratory efficiency.
978 *Cell Rep 23*, 3710-3720.
979
980 Stergiopoulos, K., Cabrero, P., Davies, S.A., and Dow, J.A. (2009). Salty dog, an SLC5
981 symporter, modulates *Drosophila* response to salt stress. *Physiol Genomics 37*, 1-11.
982
983 Wasbrough, E.R., Dorus, S., Hester, S., Howard-Murkin, J., Lilley, K., Wilkin, E.,
984 Polpitiya, A., Petritis, K., and Karr, T.L. (2010). The *Drosophila melanogaster* sperm
985 proteome-II (DmSP-II). *J Proteomics 73*, 2171-2185.
986
987 Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A.,
988 Flicek, P., Manolio, T., Hindorff, L.*, et al.* (2014). The NHGRI GWAS Catalog, a curated
989 resource of SNP-trait associations. *Nucleic Acids Res 42*, D1001-1006.
990
991 Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and
992 Romano, L.A. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol Biol*
993 *Evol 20*, 1377-1419.
994
995 Yang, P., Maguire, J.J., and Davenport, A.P. (2015). Apelin, Elabela/Toddler, and
996 biased agonists as novel therapeutic agents in the cardiovascular system. *Trends*
997 *Pharmacol Sci 36*, 560-567.
998
999 Zirin, J., Hu, Y., Liu, L., Yang-Zhou, D., Colbeth, R., Yan, D., Ewen-Campen, B., Tao,
1000 R., Vogt, E., VanNest, S.*, et al.* (2020). Large-scale transgenic *Drosophila* resource
1001 collections for loss- and gain-of-function studies. *Genetics 214*, 755-767.
1002
1003 Zirin, J., Nieuwenhuis, J., Samsonova, A., Tao, R., and Perrimon, N. (2015). Regulators
1004 of autophagosome formation in *Drosophila* muscles. *PLoS Genet 11*, e1005006.
1005

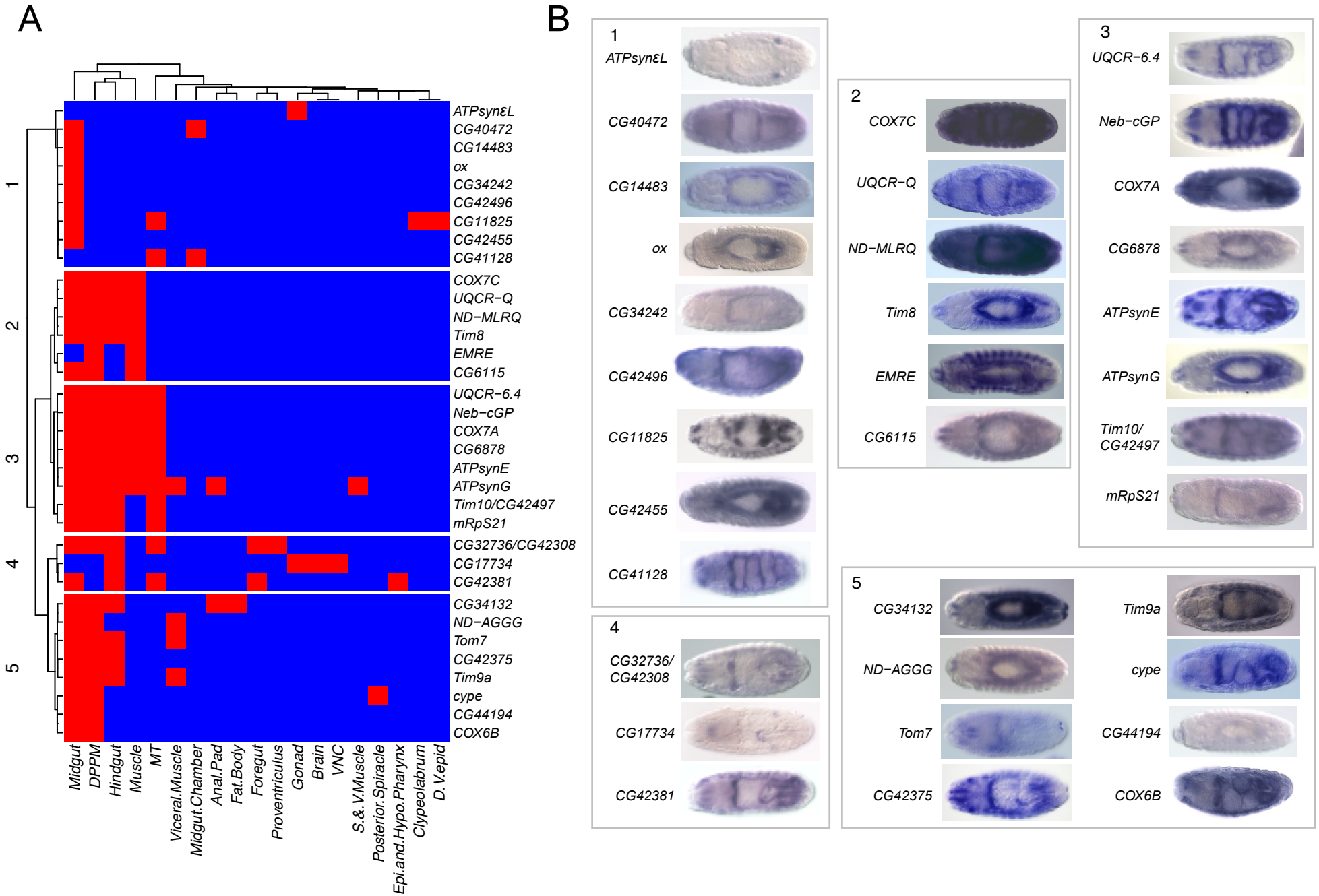
A

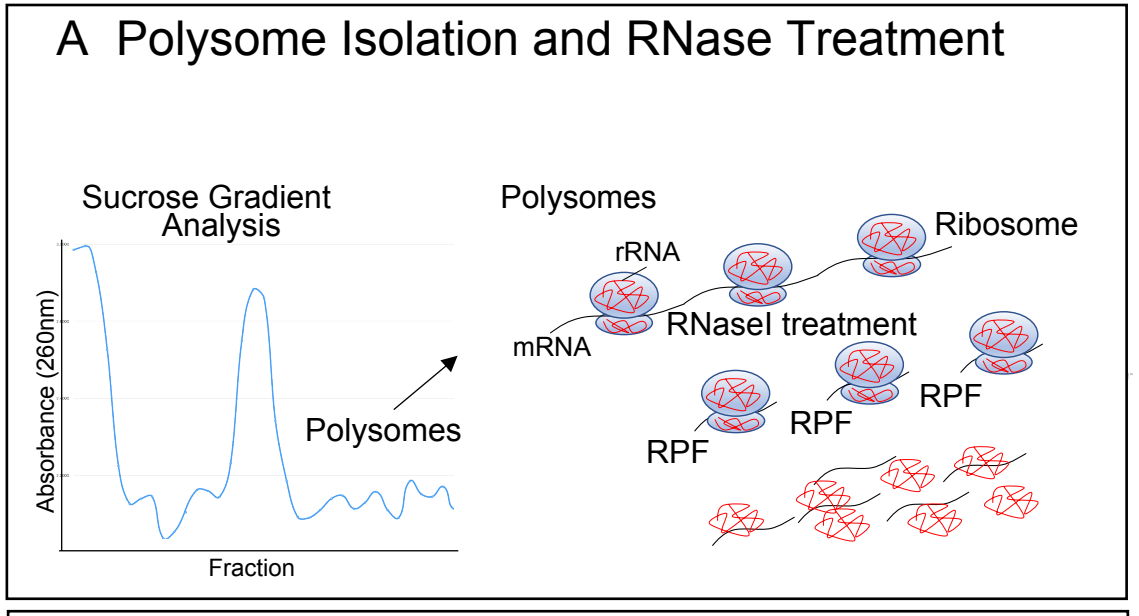

Number of smORFs
200
180
160
140
120
100
80
60
40
20
0
68
126
122
72
174
20
194
0
smORFs with Homologs
smORFs without Homologs
A. thaliana
C. elegans
D. rerio
H. sapiens/
D. melanogaster
conserved smORFs

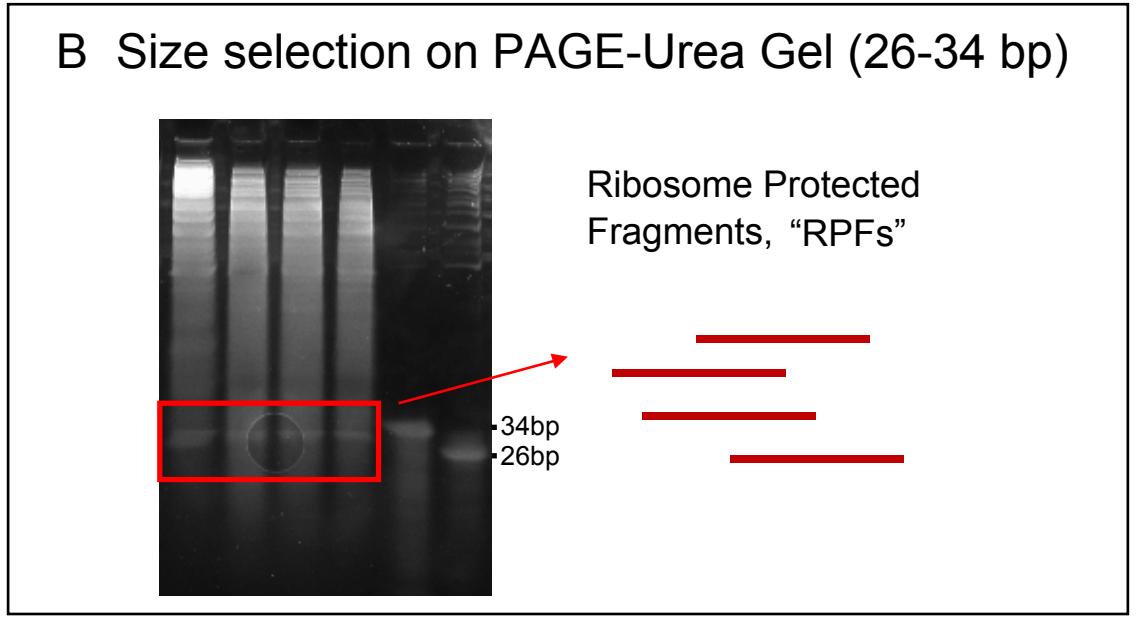
B

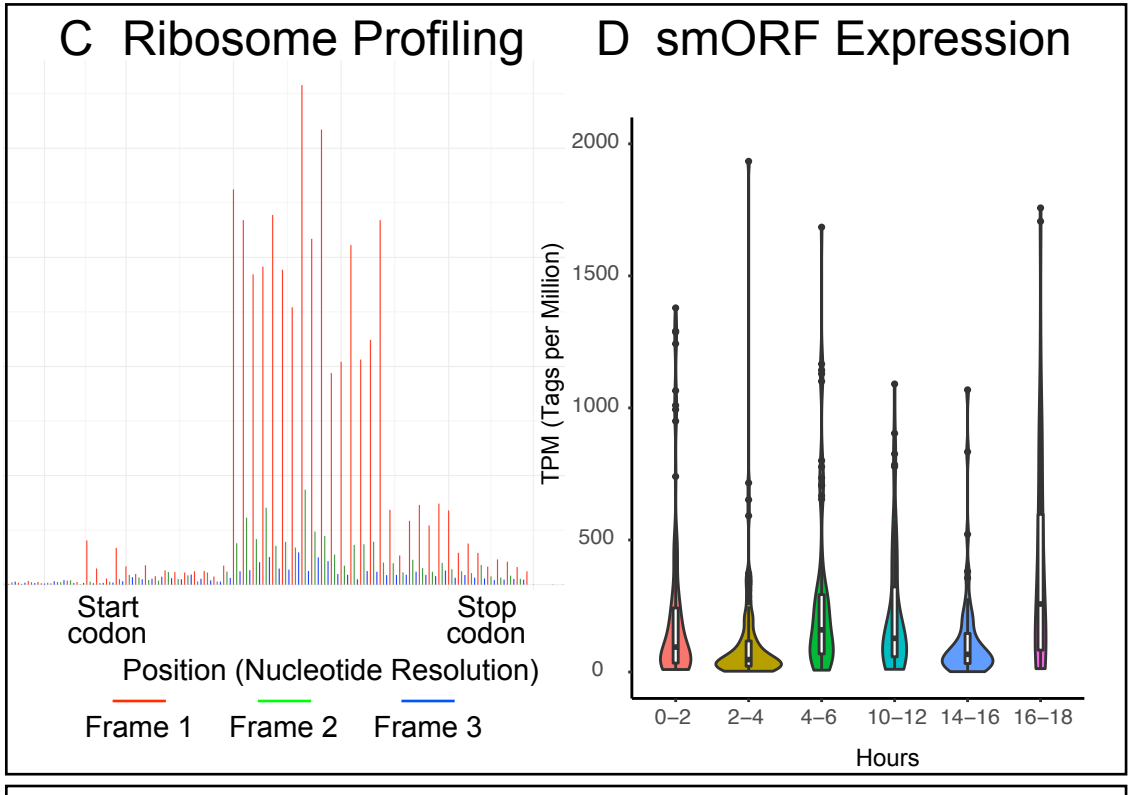

bc10
chr3L
5,565,231bp
5,566,231bp
124 insects Basewise Conservation by PhyloP
25
-4.5
Bc10-PA
Mean Hydrophobicity
Drosophila melanogaster
Anopheles gambiae
Apis mellifera
Oryzias latipes
Danio rerio
Anolis carolinensis
Gallus gallus
Mus musculus
Homo sapiens
CG42497/Tim10
chr2R
21,691,139bp
21,690,215bp
124 insects Basewise Conservation by PhyloP
25
-4.5
CG42497-PA
Tim10-PA
Mean Hydrophobicity

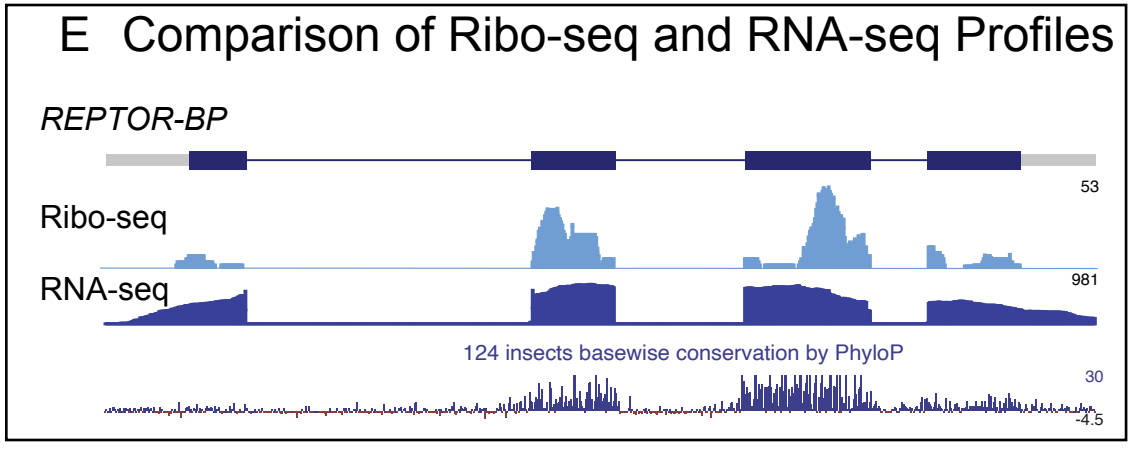

Serine-type Endopeptidase Inhibitors
Mitochondria-associated
stats
Term name
Term ID
Padj
-log10(Padj)
MF
serine-type endopeptidase inhibitor activity GO:0004867 2.91E-22
enzyme inhibitor activity GO:0004857 1.55E-19
electron transfer activity GO:0009055 4.09E-10
cytochrome-c oxidase activity GO:0004129 7.06E-10
proton transmembrane transporter activity GO:0015078 5.13E-07
BP
negative regulation of proteolysis GO:0045861 1.25E-20
negative regulation of peptidase activity GO:0010466 1.33E-20
negative regulation of endopeptidase activity GO:0010951 1.33E-20
mitochondrion organization GO:0007005 3.82E-09
cytochrome complex assembly GO:0017004 8.02E-09
electron transport chain GO:0022900 8.69E-08
respiratory chain complex IV assembly GO:0008535 2.67E-07
protein insertion into mitochondrial membrane GO:0051204 4.46E-07
ATP metabolic process GO:0046034 6.05E-07
regulation of catalytic activity GO:0050790 1.01E-06
CC
mitochondrial envelope GO:0005740 1.07E-21
mitochondrial inner membrane GO:0005743 1.13E-19
mitochondrial protein-containing complex GO:0098798 1.71E-18
mitochondrion GO:0005739 1.55E-18
mitochondrial membrane GO:0031966 3.07E-18
membrane protein complex GO:0098796 5.53E-12
cytochrome complex GO:0070069 2.79E-11
Oxidative phosphorylation KEGG:00190 3.58E-14
CG42486
CG44008
Sfp23F
Sfp24Bc
Acp24A4
CG16704
CG3513
CG16713
CG2816
CG42460
CG42464
CG42465
CG42537
CG42538
CG42713
CG42716
CG42717
CG42828
CG43350
CG15418
CG34276
CG43165
IM33
CG34423
CG14483
UQCR-6.4
UQCR-6.4L
UQCR-11L
COX7A
COX7C
cype
COX6CL
COX7AL2
ND-MLRQ
COX7CL
ATPsynE
ND-B8
CG13018
ND-AGGG
COX6AL
CG6878
Tim8
mRpL33
COX6B
UQCR-Q
ox
Neb-cGP
ATPsynG
Tom7
Tim9a
Tim13
Tim10
Fis1
CG13564
CG12479
CG41128
CG40472
CG42381
CG15458
CG17734
CG34132
CG44194
CG42496
CG34242
Cox17
ATPsynL
EMRE
CG3566
bcn92
CG11825
CG12355
CG42308
CG42375
CG42376
CG42497
mRpS21
Biological Aspect of Ancestor
Electronic Annotation
Inferred by High Throughput Experiment
Inferred from Experiment
Inferred by Curator
KEGG, Reactome


A
B
1
2
3
4
5
ATPsynεL
CG40472
CG14483
ox
CG34242
CG42496
CG11825
CG42455
CG41128
COX7C
UQCR-Q
ND-MLRQ
Tim8
EMRE
CG6115
UQCR-6.4
Neb-cGP
COX7A
CG6878
ATPsynE
ATPsynG
Tim10/CG42497
mRpS21
CG32736/CG42308
CG17734
CG42381
CG34132
ND-AGGG
Tom7
CG42375
Tim9a
cype
CG44194
COX6B
Midgut
DPPM
Hindgut
Muscle
MT
Viceral.Muscle
Midgut.Chamber
Anal.Pad
Fat.Body
Foregut
Proventriculus
Gonad
Brain
VNC
S.&.V.Muscle
Posterior.Spiracle
Epi.and.Hypo.Pharynx
Clypeolabrum
D.V.epid

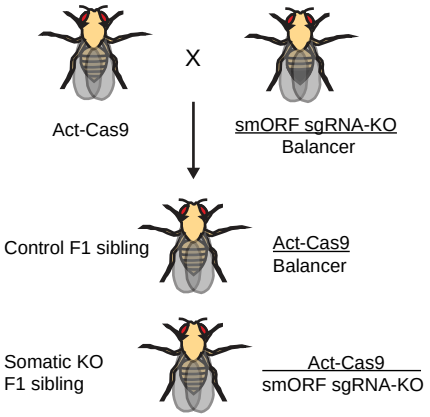

A Polysome Isolation and RNase Treatment
Sucrose Gradient Analysis
Absorbance (260nm)
Polysomes
Fraction
Polysomes
Ribosome
rRNA
mRNA
RNaseI treatment
RPF
RPF
RPF
RPF


B Size selection on PAGE-Urea Gel (26-34 bp)
Ribosome Protected Fragments, "RPFs"
34bp
26bp

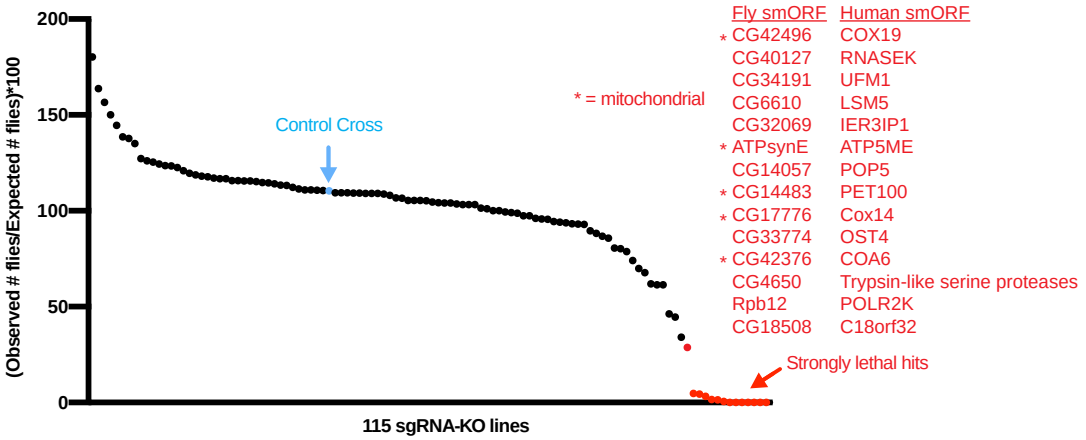

C Ribosome Profiling
Start codon
Stop codon
Position (Nucleotide Resolution)
Frame 1
Frame 2
Frame 3
D smORF Expression
TPM (Tags per Million)
2000
1500
1000
500
0
0–2
2–4
4–6
10–12
14–16
16–18
Hours


E Comparison of Ribo-seq and RNA-seq Profiles
REPTOR-BP
Ribo-seq
53
RNA-seq
981
124 insects basewise conservation by PhyloP
30
-4.5

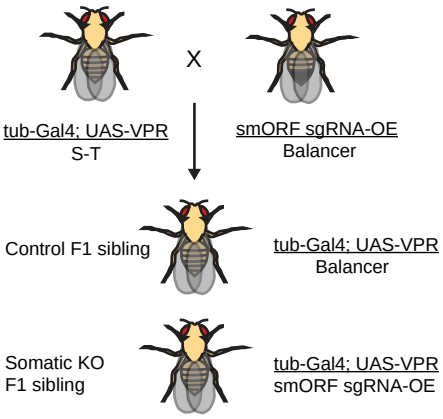
A


X
Act-Cas9
smORF sgRNA-KO
Balancer
Control F1 sibling
Act-Cas9
Balancer
Somatic KO
F1 sibling
Act-Cas9
smORF sgRNA-KO

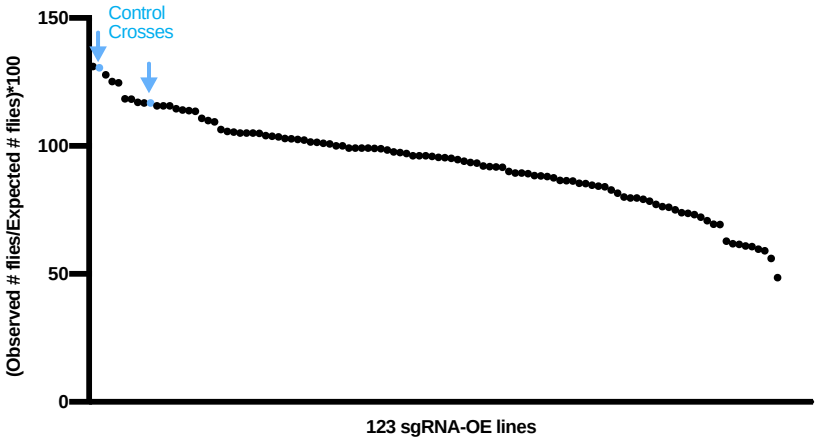
B


(Observed # flies/Expected # flies)*100
200
150
100
50
0
Control Cross
* = mitochondrial
Strongly lethal hits
115 sgRNA-KO lines
Fly smORF
Human smORF
* CG42496 COX19
CG40127 RNASEK
CG34191 UFM1
CG6610 LSM5
CG32069 IER3IP1
* ATPsynE ATP5ME
CG14057 POP5
* CG14483 PET100
* CG17776 Cox14
CG33774 OST4
* CG42376 COA6
CG4650 Trypsin-like serine proteases
Rpb12 POLR2K
CG18508 C18orf32

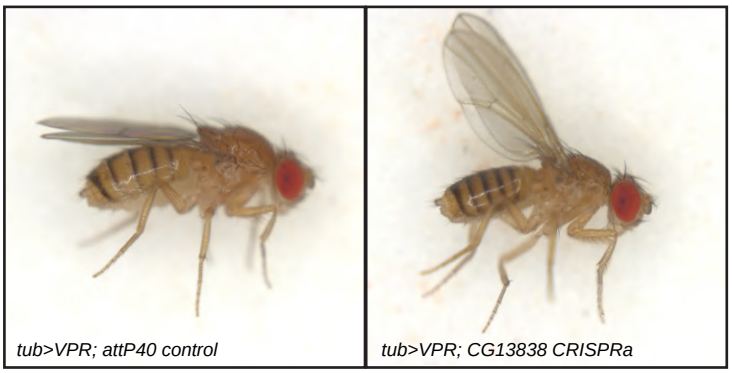
C


X
tub-Gal4; UAS-VPR
S-T
smORF sgRNA-OE
Balancer
Control F1 sibling
tub-Gal4; UAS-VPR
Balancer
Somatic KO
F1 sibling
tub-Gal4; UAS-VPR
smORF sgRNA-OE

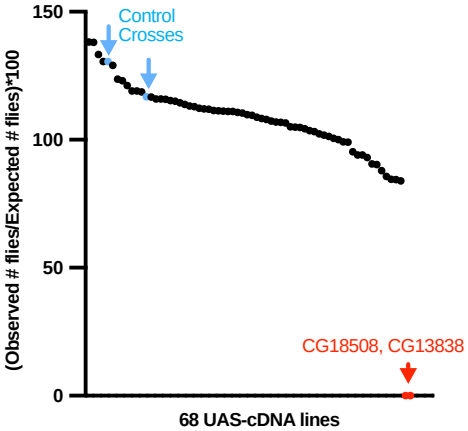
D

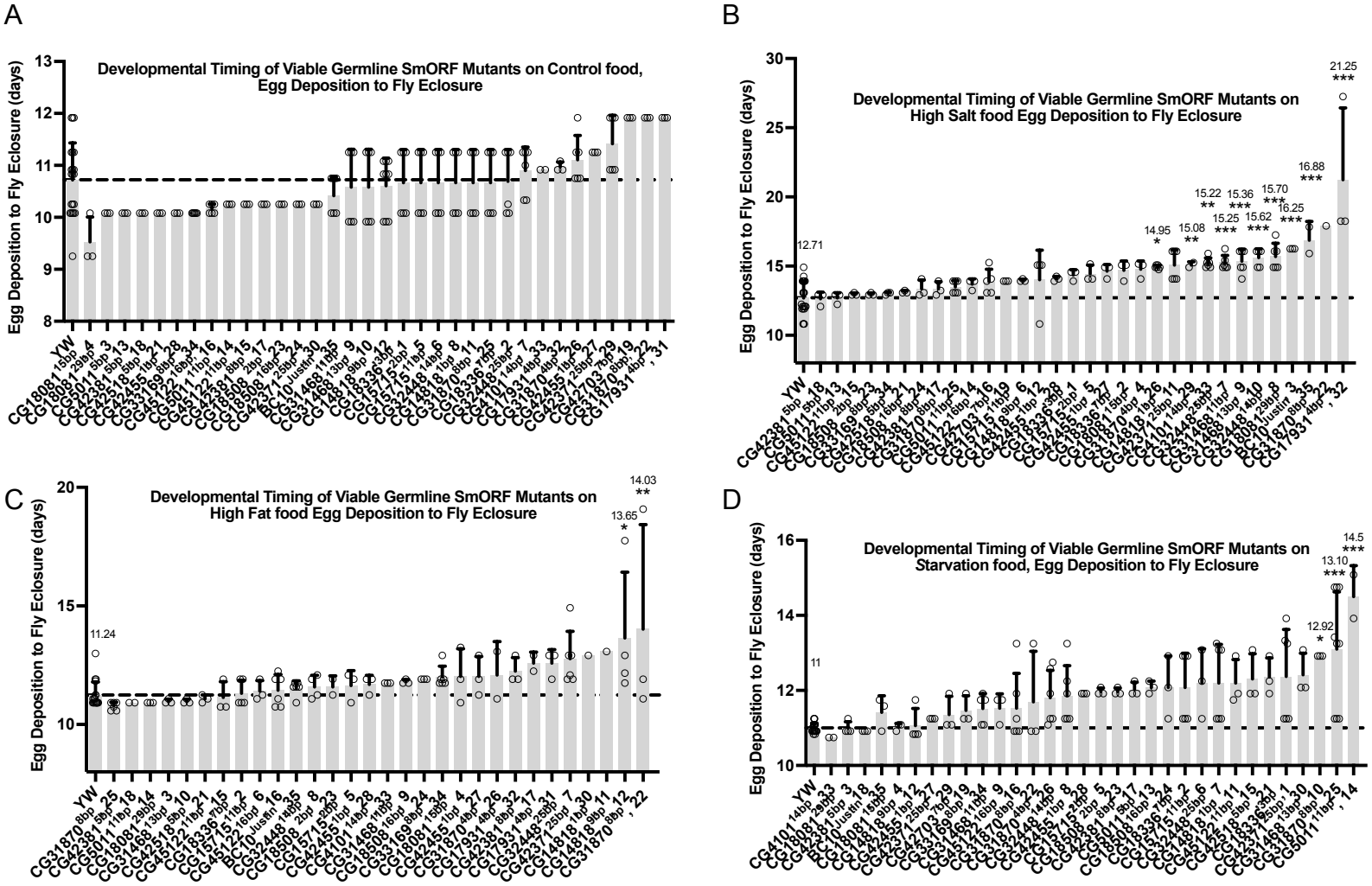

(Observed # flies/Expected # flies)*100
150
100
50
0
Control
Crosses
123 sgRNA-OE lines

E


tub>VPR; attP40 control
tub>VPR; CG13838 CRISPRa

F


(Observed # flies/Expected # flies)*100
150
100
50
0
Control
Crosses
CG18508, CG13838
68 UAS-cDNA lines


A
Developmental Timing of Viable Germline SmORF Mutants on Control food, Egg Deposition to Fly Eclosure
Egg Deposition to Fly Eclosure (days)
B
Developmental Timing of Viable Germline SmORF Mutants on High Salt food Egg Deposition to Fly Eclosure
Egg Deposition to Fly Eclosure (days)
12.71
14.95 *
15.08 **
15.22 **
15.25 ***
15.36 ***
15.62 ***
15.70 ***
16.25 ***
16.88 ***
21.25 ***
C
Developmental Timing of Viable Germline SmORF Mutants on High Fat food Egg Deposition to Fly Eclosure
Egg Deposition to Fly Eclosure (days)
11.24
13.65 *
14.03 **
D
Developmental Timing of Viable Germline SmORF Mutants on Starvation food, Egg Deposition to Fly Eclosure
Egg Deposition to Fly Eclosure (days)
11
12.92 *
13.10 ***
14.5 ***

Supplemental Figures

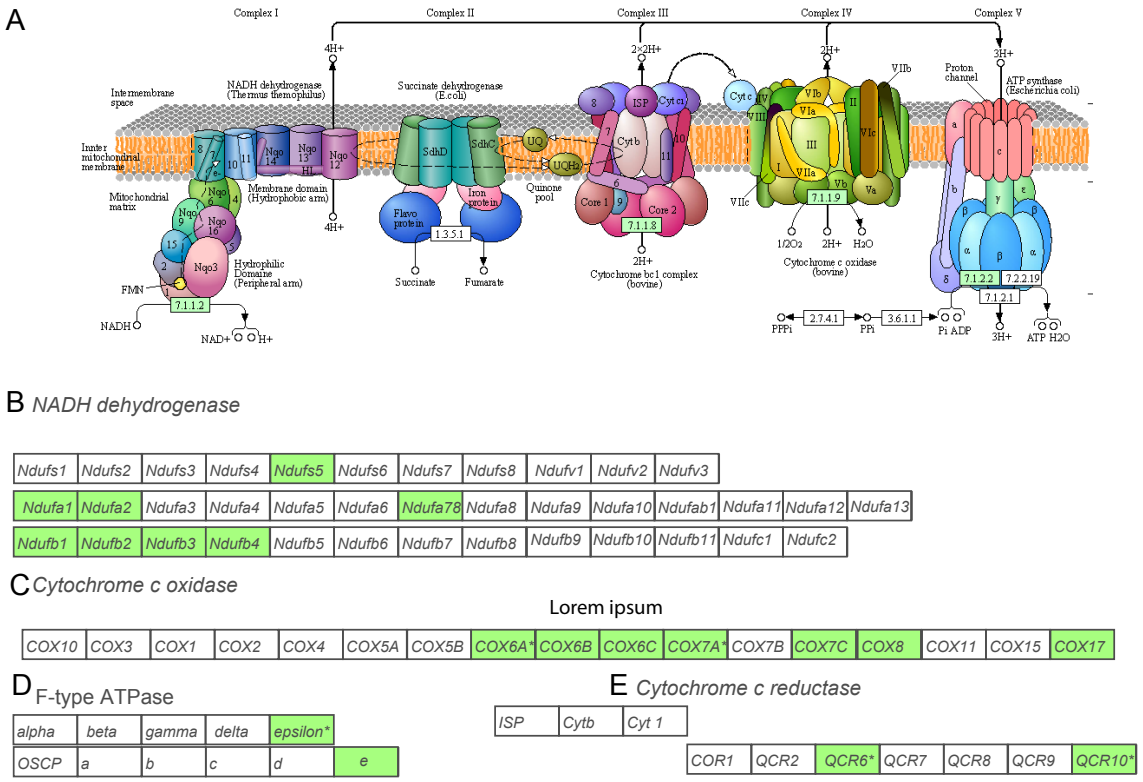


Figure S1. A) Electron transport takes place in the mitochondria and is the site of oxidative phosphorylation in eukaryotes. The complexes I to V and the enzymes required for human mitochondrial function are diagramed (Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2021; KEGG). B) Part of Complex I, NADH dehydrogenases are used in the electron transport chain for generation of ATP and to convert NADH to NAD+. Of the thirty-eight members, eight are smORFs in humans (green shading). In flies there are two orthologs for the human genes with asterisks, Ndufa7 with orthologs ND-B14.5A (103aa) and ND-B14.5AL (145aa) and Ndufb2 with orthologs CG40472 (94aa) and ND-AGGG (54aa). C) Complex IV or Cytochrome C Oxidase is a large protein complex. Of the seventeen members, seven are smORFs in humans (green shading). In flies there are two orthologs for the human genes with asterisks, *COX6A* with orthologs *levy* and *COX6AL* and *COX7A* with orthologs *COX7AL* and *COX7AL2* (62aa). There is a third ortholog of *COX6A* but it is not a smORF *COX6AL2* (289aa). D) Part of Complex V, ATPases produce ATP from ADP in the presence of a proton gradient. The human F-type ATPase episilon (ATP5F1E) has two fly orthologs *ATPsynepsilonL* (64aa) and *sun* (57aa). E) Part of Complex III, cytochrome C reductases pass electrons to Cytochrome C. Of the ten members two are smORFs and one QCR10 has two fly orthologs UQCR-6.4 (57aa) and CG43206 (72aa

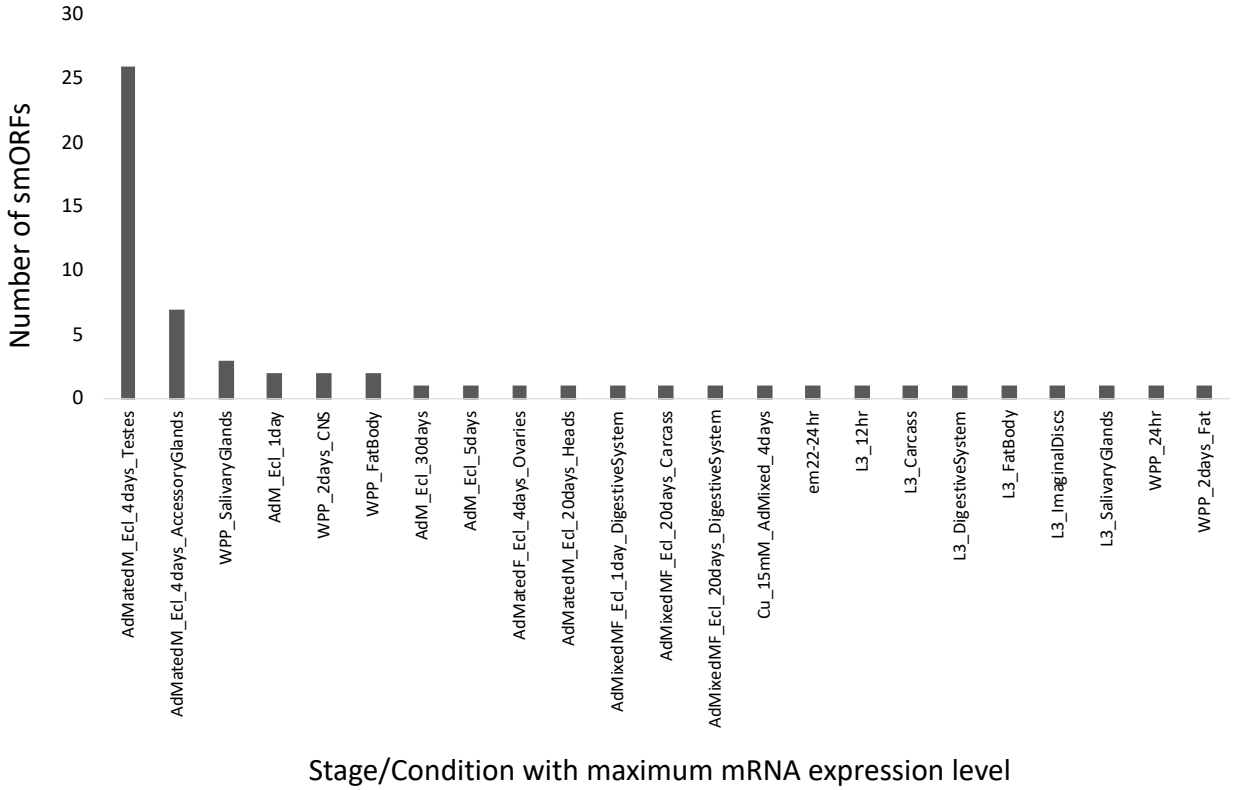


Figure S2. The developmental stage/condition of maximum mRNA gene expression (modENCODE) are plotted for ucsmORFs without Ribo-seq or proteomics support.

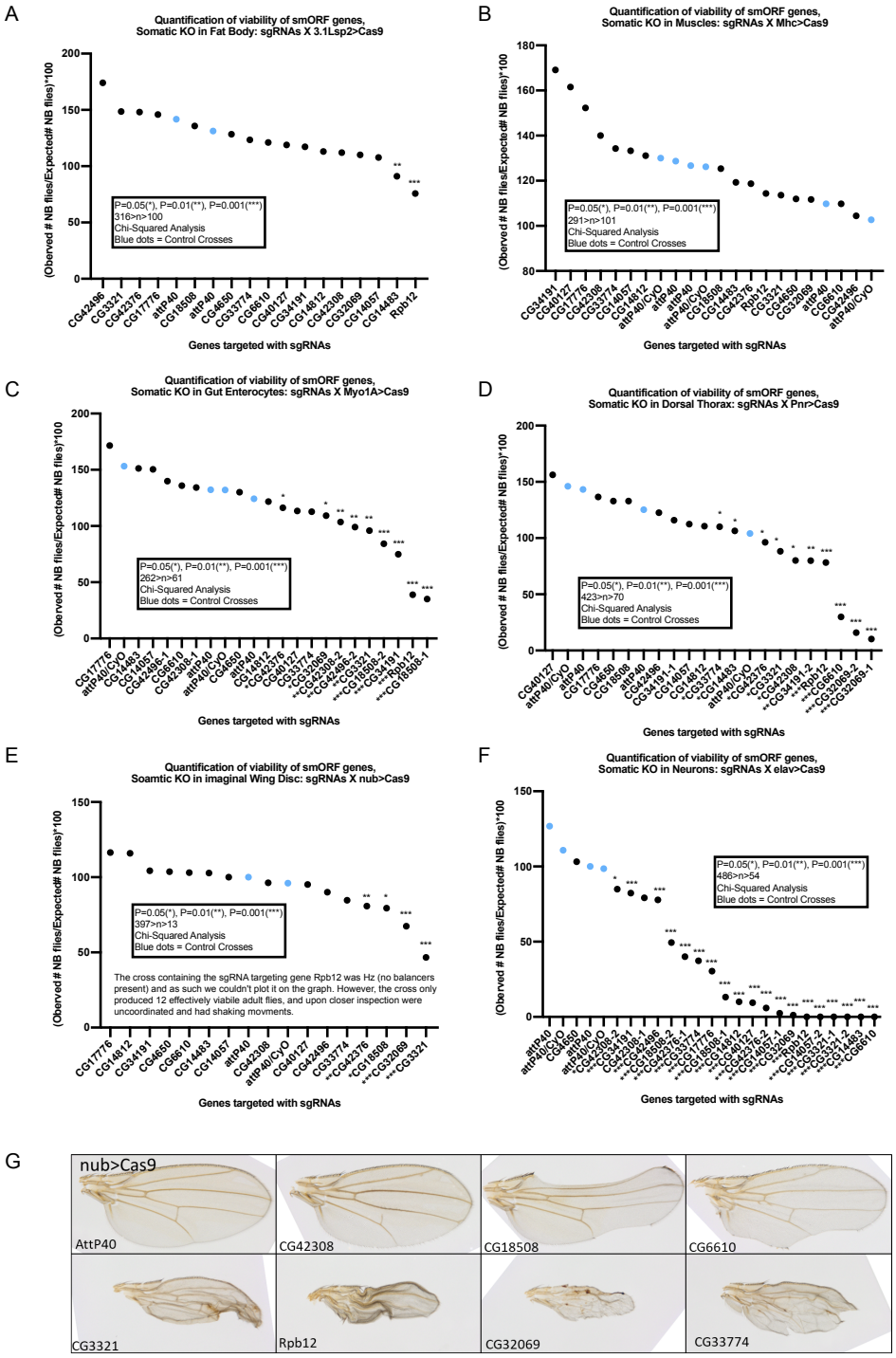


Figure S3 – Tissue-specific somatic knockout of lethal smORF hits in Figure 1B. Quantification of viability from CRISPR somatic knock out in **A)** fat body using 3.1Lsp2>Cas9, **B)** muscle using Mhc>Cas9, **C)** gut enterocytes using Myo1A>Cas9, **D)** dorsal thorax using pnr>Cas9, **E)** wing disc using nub>Cas9, or **F)** neurons using elav>Cas9. Significance calculated with a Chi-Squared Test. P=0.05(*), P=0.01(**), P=0.001(***) **G)** Images of adult wings with somatic KO.

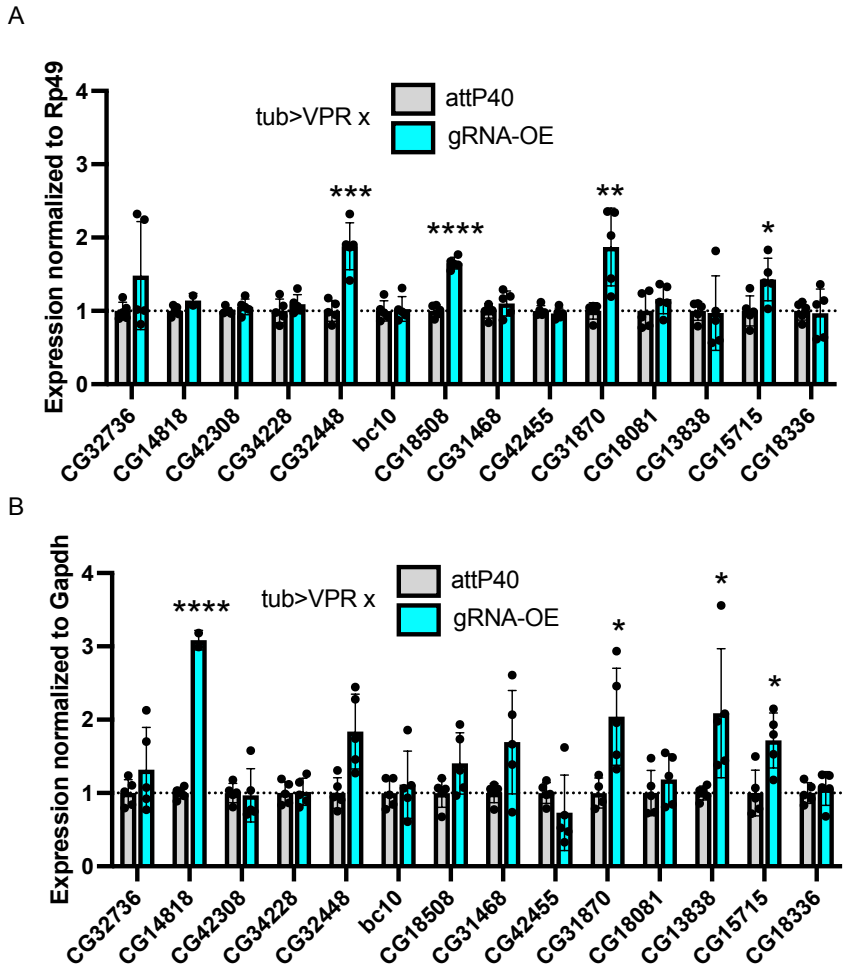
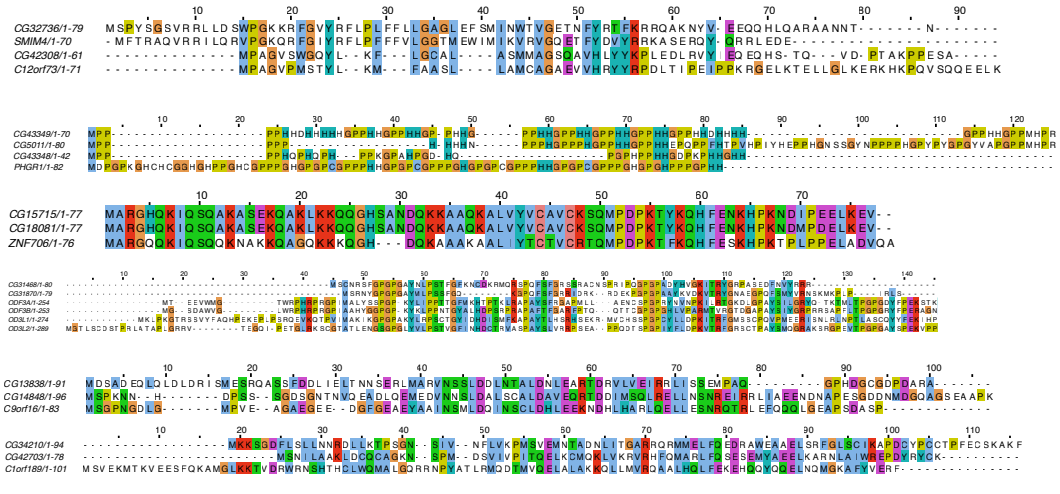


Figure S4 – qPCR verification of CRISPRa lines. smORF transcript expression normalized to (A) *Rp49* and (B) *Gapdh* for indicated genotypes. Each sgRNA-OE expression data is normalized to negative control (attP40) expression data. Significance between sgRNA-OE and attP40 control expression levels determined by a T-test. For each genotype, N=5 biological replicates, except CG14818 sgRNA-OE which had 2 biological replicates. P=0.05(*), P=0.01(**), P=0.001(***), P ≤ 0.0001 (****)

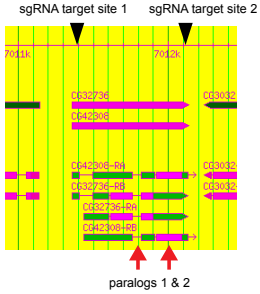
A



B

Category	smORF gene	Human Homologues	Type of KO	# alleles	Phenotype
Single gene KO	CG32736	SMIM4	Frameshift indel	1	lethal
	CG42308	C12orf73	Frameshift indel	1	lethal
	bc10	BLCAP	Knock-in gene disruption	2	viable, fertile
	CG18508	C18orf32	Frameshift indel and large deletion	4	viable, fertile
	CG33169	SMCO4	Frameshift indel	1	viable, fertile
	CG42455	SMIM13	Frameshift indel	2	viable, fertile
	CG15715	ZNF706	Frameshift indel	2	viable, fertile
	CG17931	SERF1A,SERF2,SERF1B	Frameshift indel	2	viable, fertile
	CG18081	ZNF706	Frameshift indel	1	viable, fertile
	CG18336	FAM166A	Frameshift indel	1	viable, fertile
	CG32448	SMIM8	Frameshift indel	2	viable, fertile
	CG4101	MANBAL	Frameshift indel	1	viable, fertile
	CG14818	C9orf16	Frameshift indel	2	viable, fertile
	CG31468	ODF3L1, ODF3L2	Frameshift indel	2	viable, fertile
	CG31870	ODF3L1, ODF3L2	Frameshift indel	2	viable, fertile
	CG42371	CEBPZOS	Frameshift indel	2	viable, fertile
	CG42381	CMC4	Frameshift indel	2	viable, fertile
	CG42518	NCBP2-AS2	Frameshift indel	2	viable, fertile
	CG42703	C1orf189	Frameshift indel	2	viable, fertile
	CG45122	SMKR1	Frameshift indel	2	viable, fertile
	CG5011	PHGR1	Frameshift indel	2	viable, fertile
	CG43349	PHGR1	Frameshift indel	1	viable, fertile
	CG13838	C9orf16	Large deletion	3	viable, fertile
	CG34210	C1orf189	Large deletion	3	viable, fertile
	CG34228	SMIM12	Large deletion	3	viable, fertile
Multi gene KO	CG32736/CG42308	SMIM4/C12orf73	Large deletion	1	lethal
	CG43349/CG5011/CG43348	PHGR1	Large deletion	1	viable, fertile
	CG15715/CG18081	ZNF706	Large deletion	1	viable, fertile
	CG31468/CG31870	ODF3L1, ODF3L2	Frameshift indels	1 combo	viable, fertile
	CG13838/CG14818	C9orf16	Frameshift indel and Large deletion	1 combo	viable, fertile
	CG34210/CG42703	C1orf189	Large deletion and frameshift indel	1 combo	TBD

C



D

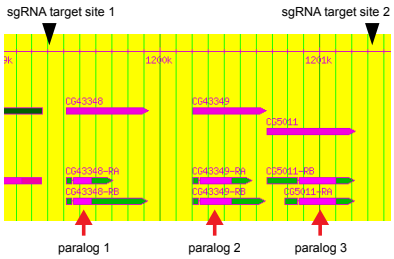


Figure S5 – Full CRISPR knockout of 25 uncharacterized smORF genes and muli-gene knockout of paralogs. (*A*) Protein sequence alignment of smORF paralogs listed in Figure 2B with their human homologs. **(B)** List of 25 uncharacterized smORFs, their human homologs and fly paralogs, type of molecular knockout, and homozygous fly phenotype. **(C, D)** Genomic views paralogs with location of sgRNA target sites for multi-paralog deletion.