Title
Molecular and functional characterization of the Drosophila melanogaster conserved smORFome.

Journal
Cell Reports, 42(11)

Authors
Bosch, Justin
Keith, Nathan
Escobedo, Felipe
et al.

Publication Date
2023-11-28

DOI
10.1016/j.celrep.2023.113311

HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2024 February 05.

Published in final edited form as:

Cell Rep. 2023 November 28; 42(11): 113311. doi:10.1016/j.celrep.2023.113311.

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

Justin A. Bosch[1,6], **Nathan Keith**[2,3,6], **Felipe Escobedo**[1], **William W. 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]Senior authors

[8]Lead Contact

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 298 smORFs that are evolutionarily conserved between *Drosophila melanogaster* and humans. Many of these smORFs are conserved broadly in the bilaterian lineage, with ~182 conserved in plants. Within these conserved smORFs, we observe remarkably heterogenous spatial and temporal expression patterns – indicating widespread tissue-specific and stage-specific mitochondrial architectures. In addition, an analysis of annotated functional domains reveals a predicted enrichment of smORF polypeptides localizing to mitochondria. We conduct an embryonic ribosome profiling experiment finding support for translation of 137 of these smORFs during embryogenesis. We further embark on functional characterization using CRISPR knockout/activation, RNAi knockdown, and

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

Author Contributions:

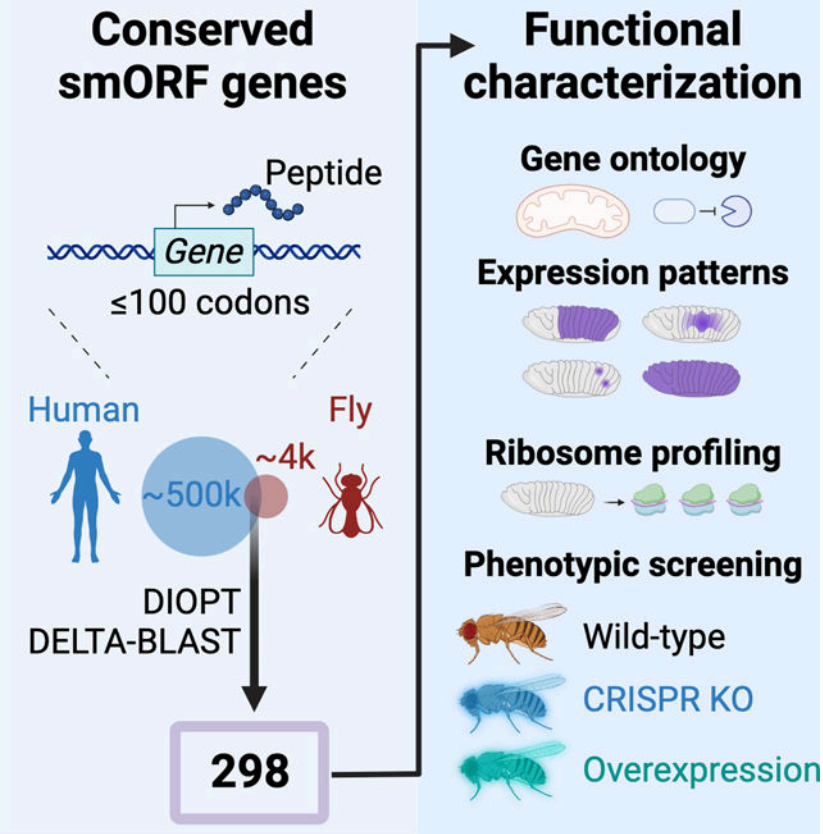
J.A.B and N.K. designed and performed experiments, generated reagents, and wrote the manuscript. F.E., W.W.F., J.T.L., J.R., K.H.W., and R.W. performed experiments and generated reagents. Y.H. performed data analysis. S.K. provided CG42371 KO flies. J.B.B., N.P., and S.E.C. supervised, designed experiments, and edited the manuscript.

Declaration of Interests:

The authors declare no competing interests.

cDNA overexpression, revealing diverse phenotypes. This study underscores the importance of identifying smORF function in disease and phenotypic diversity.

Graphical Abstract



Introduction:

Genome annotations have often overlooked proteins with less than 100 amino acids, although many have been shown to play important roles in development and physiology and are pervasive across the Tree of Life[1–3]. While some small proteins are cleavage products of longer proteins, many others are encoded in the genome by small open reading frames (smORF genes; δ 100 amino acids). Strikingly, it has been reported that human disease-associated variants from genome-wide association studies (GWAS) are enriched in smORF genes[4]. These estimates underscore the functional roles of smORFs and relevance to human diseases.

Advances in proteomics and next-generation sequencing (NGS) technologies have led to a significantly improved annotation of smORF genes. There is evidence for >2,500 smORF genes in humans[5], and there are over 1,000 annotated smORF genes in the *Drosophila melanogaster* genome[6]. Some smORF genes are important regulators of physiology, development and metabolism and encode hormones[7], neurotransmitters[8], ligands and cofactors[9], RNA and DNA binding factors, and components of ribonucleoproteins[10].

Interestingly, a number of them are involved in numerous mitochondrial functions and processes[11–13]. Notably, studies in insects have defined the functions of several smORF peptides, such as Tarsal-less/mille-pattes/polished-rice[14–17], Brd[18–20], and Pgc[21].

Although smORF genes are prevalent in metazoan genomes, a surprisingly small number of these genes are evolutionarily conserved in animals suggesting a high birth and death rate for these genes[5,22]. 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[5]. 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. Again, studies in Drosophila have pioneered the bioinformatic identification of smORF peptides, using techniques such as amino acid conservation, ribosomal profiling, and proteomics[23–26].

Here, we characterize a collection of 298 fly smORFs conserved with human. 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

We identified 298 smORF genes that are evolutionarily conserved between humans and *Drosophila melanogaster*, which we refer to as conserved smORFs (Figure 1A). Briefly, we used two strategies, DIOPT and DELTA-BLAST, to identify human-fly orthologs (Materials and Methods; Supplemental File 1). All 298 are currently annotated as protein coding. We further analyzed conservation of these smORFs with the well-annotated transcriptomes of zebrafish (*Danio rerio*), nematodes (*Caenorhabditis elegans)* and thale cress (*Arabidopsis thaliana)* (Figure 1B). There are 274 conserved in zebrafish, and 239 conserved in *C. elegans*. Notably, 182 conserved smORFs were also conserved in *Arabidopsis*, providing evidence for the functional importance of this dataset in non-Bilateria eukaryotes. Finally, amino acid alignment of many of the human-fly conserved smORFs, such as *bc10*, *CG42497*, and *Tim10,* reveal that they are also conserved among other invertebrate and vertebrate species (Figure 1C).

Of the 298 *Drosophila* conserved smORFs, 32 are polycistronic (Supplemental File 1, see example in Figure 1C). Remarkably, while the individual smORFs that reside in *Drosophila* polycistronic transcripts are evolutionarily conserved, their polycistronic structure is generally not - indicating a complex evolutionary history. Interestingly, there are three conserved smORFs encoded by polycistronic transcripts in both fly and *C. elegans*, *CG42372, CG42375* and *Mocs2A*. Between flies and zebrafish, there are two smORFs encoded by polycistronic transcripts in both species, *CG42497* and *Mocs2A*. However,

between flies and humans, or flies and *Arabidopsis,* there are none. All non-fly smORF orthologs are currently annotated as protein-coding.

Gene Ontology analysis of conserved smORFs

The functions of conserved smORFs are diverse, as with any broad category of genes. Gene Ontology (GO) Cellular Component enrichment analysis of conserved smORFs determined that the majority of significantly enriched GO terms are associated with mitochondrial function and localization (Figure 2, Supplemental File 2). Indeed, 66 conserved smORFs are predicted to be involved in mitochondrial function (Supplemental File 2). "Mitochondrion" ($P = 1.46 \times 10^{-51}$) which contains 63 conserved smORFs is the most significantly enriched GO Cellular Component terms in this dataset, with "Mitochondrial Envelope" as the second most significantly enriched GO Cellular Component term ($P = 2.2 \times 10^{-49}$) (Figure 2). Additional significantly enriched terms include mitochondrial inner membrane ($P = 1.4 \times 10^{-45}$) and cytochrome complex ($P = 6.09 \times 10^{-24}$) (Figure 2).

The oxidative phosphorylation pathway is the only significantly enriched pathway in the smORF dataset ($P = 7.33 \times 10^{-31}$;[27] KEGG; Supplemental File 2, Figure 2, Figure S1). Four of the genes in the oxidative phosphorylation pathway, *COX6CL*, *CG40472, COX7CL,* and *UQCR-6.4L,* are paralogs of annotated fly genes: *cyclope* (*cype*) encoding the *cytochrome c oxidase subunit 6C*, *NADH dehydrogenase (ubiquinone) AGGG subunit* (*ND-AGGG*), *Ubiquinol-cytochrome c reductase 6.4 kDa subunit* (*UQCR-6.4*) and *Cytochrome c oxidase subunit 7C* (*COX7C*). Interestingly, *COX6CL, COX7CL and UQCR-6.4L* are primarily expressed in the adult testis, whereas their paralogs (*cype*, *COX7C,* and *UQCR-6.4*, respectively*)* are ubiquitously expressed[28].

In contrast to the GO Cellular Component enrichment analysis, the most significantly enriched Biological Function and Molecular Process GO terms are related to serine-endopeptidase inhibitor activity (Figure 2; Supplemental File 2). For instance, the most significantly enriched Molecular Function GO term is serine-type endopeptidase inhibitor activity ($P = 2.91 \times 10^{-22}$) (Figure 2). Notably, 32 of the 80 predicted serine-endopeptidase inhibitors in the *D. melanogaster* genome are in the conserved smORF dataset. Of these 32 smORFs, 22 contain a predicted pancreatic trypsin inhibitor Kunitz domain (Interpro: IPR036880), and 10 smORFs contain a predicted a Kazal domain (Interpro: IPR036058) - providing evidence that these 32 conserved smORFs are serine-endopeptidase inhibitors[29]. Additionally, all 32 of these predicted serine-endopeptidase inhibitors contain an N-terminal secretion signal (SignalP6.0; 0.982 or higher score)[30].

In situ imaging of smORF mRNA expression in embryos

To better understand the functional roles of smORFs in *Drosophila*, we performed in situ mRNA hybridization to visualize smORF expression during embryogenesis (Supplemental File 3; https://insitu.fruitfly.org/). Of these, organ-specific expression (i.e., "patterned" expression) could be assigned for 143 conserved smORFs during embryonic development (Supplemental File 3).

In addition to patterned expression, conserved smORFs can also be classified as being "maternally deposited", and therefore ubiquitously expressed in the embryo in the

earliest developmental stage(s), and/or classified as being “ubiquitously expressed”, where expression is observed throughout the full embryo after the earliest developmental stage (Supplemental File 4). We found that 59 conserved smORFs were classified as only being maternally deposited and/or being ubiquitously expressed (i.e., expressed throughout the entire embryo without assigned organ patterns) in at least one embryonic stage. Of these 59 smORFs, *CG15456* and *UQCR-6.4L* are the only smORF with no observed expression after maternal deposition (i.e., after initial embryonic stages 1–3).

The remaining 86 smORFs revealed no observed expression in the embryo. In each case this was concordant with RNA-seq[31]: these smORFs are expressed later in development, and in specific tissues, or under stress conditions[31].

In situ imaging of mitochondria-associated smORF mRNA expression

Notably, mitochondria-associated conserved smORFs exhibited heterogenous spatial expression patterns, indicating tissue- and stage-specific mitochondrial architectures, and, by extension, tissue- and stage-specific mitochondrial functions (Figure 3). Of these, expression could be assigned as patterned for 42 smORFs. (Figure 3). We clustered these mitochondria-associated conserved smORFs with patterned expression into six groups - showing extensive inter- and intra-group heterogeneity in organ system expression patterns (Figure 3).

Nearly all (42/45) patterned mitochondria-associated conserved smORFs were maternally deposited (Supplemental Files 3 and 4), and 32/34 were annotated as being ubiquitously expressed in at least one later embryonic stage (Supplementary File 3). Of the two conserved smORFs without ubiquitous expression after maternal deposition, one (*Cox17*) shows robust embryonic expression under RNA-seq[28], but *in situ* hybridization was unsuccessful; while the other is only annotated as patterned post maternal deposition (*CG17734*) (Supplemental File 4). However, nearly all patterned, mitochondrial smORFs (33/34) exhibit complex and tissue specific expression patterns after early embryonic ubiquitous expression (usually due to maternal deposition) (Supplemental File 4). Hence, the zygotic regulation of structural and functional components of mitochondria is highly patterned and specific to individual tissues and cell types.

Ribosome profiling and proteomics analysis to support the translation of conserved smORFs

To verify the translation of conserved smORFs, we performed ribosome profiling for six, 2-hour embryonic stages and six, 0–24 hr. mixed stage embryo samples (Figure 4; Figure S2, Supplemental Files 5). We find evidence of translation for 137 (46%) conserved smORFs from these embryonic samples. We find that 42 (14%) conserved smORFs are detected in only a single embryonic stage (Supplemental File 5), consistent with transcriptional evidence from extensive RNA-seq experiments in a developmental time course[28,32].

We analyzed previous ribosomal profiling studies in *Drosophila* embryos[33,34], which provided translation evidence for an additional 107 conserved smORFs (Supplemental File 5). Nine smORFs (*CG13784*, *CG42496*, *Atg8a*, *Ccdc56*, *CG15386*, *CG31313*, *glob1*, *MED18*, *Pis*) were uniquely identified in our study (Supplemental File 5). Similarly, we analyzed published mass spectrometry datasets[24,31,35–40], which gave polypeptide support

for 186 total conserved smORFs (Supplemental File 5). All datasets combined provide support for a total of 254 (85%) conserved smORFs, leaving 44 with no translation or polypeptide support. Analysis of comprehensive *Drosophila* modENCODE mRNA expression data[28] revealed 24/44 (55%) of the conserved smORFs without evidence for translation or peptides are maximally expressed in testes or accessory gland (Figure S3). Notably, 23/24 (96%) of these conserved smORFs that are maximally expressed in testes and accessory gland are classified as having no or low mRNA expression throughout embryogenesis, and their expression is testes- and/or accessory gland-specific[28,32]. Interestingly, 17/24 (71%) of these smORFs have a human homolog that are expressed in testes (GTEx Portal, TPM >1).

Functional analysis of smORF genes by F1 CRISPR screens

Next, we assessed if conserved smORFs have important biological functions in *Drosophila* by modifying their gene function in vivo. First, to knock out (KO) smORF gene function in a systematic manner, we used a CRISPR/Cas9-based transgenic crossing strategy[41,42] where a Cas9-expressing line is crossed with sgRNA lines that target 5' coding sequence (sgRNA-KO). The resulting progeny will contain somatic indels in the target gene that disrupt gene function. We generated a collection of 177 sgRNA-KO lines that target 165 smORF genes (Supplemental File 7). Each smORF sgRNA-KO line was crossed with *Act5c-Cas9* (ubiquitous Cas9) and F1 progeny were screened for defects in viability, morphology, or gross motor behavior (Figure 5A). Of the 115 sgRNA-KO lines tested, 14 (representing 14 genes) gave no mutant adult progeny or very few compared to controls (Figure 5B; Supplemental File 7), suggesting that they are essential genes. No other obvious morphological or behavioral phenotypes were observed for the remaining sgRNA-KO lines.

To determine if the 14 putative essential smORFs play an important role in individual tissues, we crossed the 14 sgRNA-KO hit lines to cell-type specific Cas9 lines (muscles, gut enterocytes, dorsal thorax, wing disc, neurons) (Figure S4). Nearly all sgRNA-KOs reduced viability when expressing Cas9 in neurons (12/14), whereas none reduced viability with muscle Cas9 (0/14). Interestingly, a subset of sgRNA-KO lines had reduced viability only when Cas9 was expressed in neurons (CG14057, CG40127, CG14812, CG17776). In contrast, Rbp12 sgRNA-KO was lethal or showed low viability with all Cas9 lines except muscle, and CG4650 sgRNA-KO did not cause any obvious phenotypes with any of the six tissue-specific Cas9 lines. Finally, using a larval wing disc Cas9 line, seven sgRNA-KO lines caused adult wing defects, such as notching or crumpled wings (Figure S4).

To over-express smORF genes in a systematic manner, we used CRISPR activation (CRISPRa), where a sgRNA targets the promoter region and increases expression of the endogenous gene (sgRNA-OE) via a catalytically dead Cas9 (dCas9) fused with a transcriptional activator (e.g. VPR or SAM)[43,44]. Similar to our sgRNA-KO collection, we generated a collection of 197 transgenic sgRNA-OE lines that target 176 smORF genes. Each smORF sgRNA-OE line was crossed with *tub-Gal4*, *UAS-dCas9-VPR* (abbreviated *tub>VPR*), and the F1 progeny were screened for phenotypes as described for the sgRNA-KO collection (Figure 5C; Supplemental File 7). Of the 123 sgRNA-OE lines tested, a small number had reduced numbers of expected progeny, however none were statistically

significant when compared to control crosses (Figure 5D). Interestingly, CG13838 sgRNA-OE resulted in viable adults that were flightless and had a “held-up” wing phenotype (Figure 5E). Both phenotypes were 100% penetrant (n=100 flies). None of the remaining tested lines produced aberrant morphological or behavioral phenotypes. To validate overexpression by CRISPRa, we crossed 14 smORF-OE lines to *tub>VPR* and analyzed target gene expression in adults using quantitative PCR (qPCR) (Figure S5). These results showed that 6/14 sgRNA-OE lines had significantly elevated transcript expression when normalized to *Rp49* or *Gapdh* and compared to negative control crosses (*attP40*).

To complement our results with CRISPRa, we overexpressed a subset of smORFs using *UAS-cDNA* lines (Figure 5F; Supplemental File 7), which generally results in higher levels of overexpression[43]. Of the 63 lines tested, representing 58 genes, two lines were lethal, *UAS-CG18508* and *UAS-CG13838*. All other lines had no obvious defects in viability, morphology, or behavior.

Functional analysis of 25 uncharacterized smORF genes by whole animal KO

F1 CRISPR screening tools are fast and scalable but have technical limitations, namely that CRISPR-KO can produce mosaic phenotypes[42] or phenotypes outside the target tissue[45]. Therefore, we wanted to apply a more robust genetic tool to modify smORF gene function – whole animal knockout. Since this requires greater time and resources, we targeted a subset of the conserved smORFs.

Using a combination of gene function prediction and manual searching (see Materials and Methods), we identified 25 conserved smORF genes with minimal to no previous experimental characterization in any organism with a corresponding homolog (Supplemental File 7). Interestingly, 12 of these smORFs have a paralog in *Drosophila* (Figure S6A). Using CRISPR/Cas9, we generated whole animal KOs for each of the 25 smORF genes and multi-gene KOs for each paralog group (Supplemental File 7; Figure S6B–D). When possible, we generated at least two independently derived alleles for each smORF. The resulting homozygous animals were assessed for mutant phenotypes. Remarkably, nearly all smORF KO lines were viable, fertile, and had no obvious morphological or behavioral defects (Supplemental File 7). Interestingly, KO of the paralogs *CG32736* and *CG42308* was lethal, either as single or double KO (Supplemental File 7; Figure S6B–C). Characterization of *CG32736* and *CG42308* is described elsewhere[11].

We reasoned that viable smORF KO mutants might reveal a mutant phenotype if raised under stressful conditions. To test this, we transferred 24hr old homozygous smORF KO embryos onto modified foods known to cause animal metabolic stress, starvation[46], high fat[47], and high salt[48], and measured developmental timing. On normal food, all tested KO mutants had similar developmental timing compared to wild-type (Figure 6A). However, several KO mutants exhibited significant developmental delays, low viability, or lethality on stressful foods (Figure 6B–D). For example, two independent CG17931-KO alleles were lethal on starvation food and were developmentally delayed or low viability on high salt food. In addition, two independent alleles of CG42371-KO showed low viability on high salt food, and one CG42371 allele had low viability on high fat food. Finally, bc10-KO was developmentally delayed on high salt food.

Discussion

We identified 298 conserved smORFs between humans and fruit flies, the vast majority of which are conserved broadly across bilaterians, with 68 conserved in *Arabidopsis*. The decreased number of smORFs in C.elegans, zebrafish, and Arabidopsis may be due to gene loss or extreme divergence. Ribosome profiling and prior proteomic experiments support the translation of 208 of these smORFs. The remaining 90 without direct evidence of translation are largely highly tissue specific (50 are specific to male reproductive tissues), and likely more targeted samples are needed for detection. Of course, the thresholds we selected for defining smORFs are somewhat arbitrary, and more lenient parameterizations (>100aa) would yield more expansive lists. Certainly, this threshold produced an under-studied collection of genes with diverse functions and phenotypes.

Of the 298 conserved smORFs, 32 are arranged in polycistronic transcripts. However, this gene architecture is rarely conserved in distant species, with only four remaining polycistronic in distant species – the remainder are partitioned into distinct transcriptional units, often on different chromosomes.

The largest class of conserved smORFs is related to mitochondrial structure and function – including components of the oxidative phosphorylation pathway. Two of these predicted mitochondrial smORFs were recently functionally characterized[11], though it possible that not all function in mitochondria. For example, 22 out of 66 predicted mitochondrial smORF genes are uncharacterized “CGs”.

Similarly, human mitochondria are enriched in smORF peptides[49]. Intriguingly, *Drosophila* mitochondrial smORFs exhibit highly tissue specific expression patterns after initial ubiquitous maternal deposition. While mitochondrial functional diversity has been explored in the nervous system[50], this study indicates a far broader diversity in mitochondrial architecture throughout the developing organism. The profound conservation of these genes, along with their tissue-specific expression patterns, indicates that mitochondria are compositionally, and therefore functionally, optimized in a tissue-specific fashion. This observation points to an evolutionary impetus for the translocation of mitochondrial genes to the host nuclear genome -- tissue-specific regulation by host nuclear factors – an intriguing direction for future study.

Results from our F1 CRISPR knockout screens revealed a number of essential smORF genes. Interestingly, animal homologs of these gene hits (10/14) show lethality in other organisms (Marrvel.org)[51], and 5/14 are predicted mitochondrial (Supplemental File 2), suggesting that fly lethality may be due to disrupted mitochondrial function. Furthermore, these essential smORFs may be required in different tissues. For example, Rbp12 knockout in either the gut, dorsal thorax, wing disc, or neurons caused significant reduction in animal viability. In contrast, four essential smORFs were only lethal when knocked out in neurons.

One interesting hit from our F1 knockout screen was *CG18508,* the fly homolog of *C18orf32.* The encoded protein has been shown to associate with lipid droplets[52]. While ubiquitous *CG18508* knockout reduced viability, knockout in six tissues did not. However, *CG18508* knockout in the wing disc caused adult wing notching. Surprisingly, homozygous

CG18508 mutants are viable, fertile, and had normal wing morphology, suggesting that *CG18508* sgRNA-KO has off-target effects. Interestingly, overexpression of *CG18508* by UAS-cDNA was lethal. While it is not clear if *CG18508* overexpression if physiologically relevant, it would be intriguing to examine if animals die due to defects in lipid storage.

Results from the smORF overexpression screen revealed one additional smORF with notable phenotypes. Overexpression of *CG13838* by CRISPRa resulted in flightless adult flies with "held-up" wings, whereas overexpression by UAS-cDNA was lethal. Since UAS-cDNA generally results in higher transcript expression compared to CRISPRa[43], *UAS-CG13838* cDNA lethality may be due to higher expression. The *C. elegans* homolog of *CG13838*, *bubblin* (*bbln*), has recently been characterized as essential for intermediate filament function[53]. Interestingly, other fly mutations have been described that cause a wing phenotype, such as *heldup*, which disrupts muscle thin filaments[54]. Like intermediate filaments, thin filaments are actin-based cytoskeletal structures[55]. Therefore, *CG13838* overexpression may interfere with thin filaments and/or intermediate filaments in muscle.

Many conserved smORFs have known or predicted functions. For example, smORF CG14483 is uncharacterized in *Drosophila*, but its human homolog PET100 (65% aa similarity, 37% aa identity) is a known regulator of mitochondrial complex IV biogenesis and mutated in families with mitochondrial complex IV deficiency nuclear type 12 (MC4DN12)[56]. In contrast, we identified 25 smORFs with little to no characterization in any organism. For example, CG32736 is homologous to human Small Integral Membrane Protein 4 (SMIM4) (66% aa similarity, 46% aa identity), but had not been experimentally studied in any organism until recently[11,57,58]. Studying poorly characterized smORF genes like *CG32736* could reveal new biology and/or help understand human disease progression. Interestingly, we found three previously uncharacterized smORFs (*CG17931/SERFs*, *CG42371/CEBPZOS*, *bc10/BLCAP*) that were required for normal developmental progression on stressful food diets.

Limitations of the Study:

Our list of 298 fly-human conserved smORFs is likely incomplete. Indeed, the fly smORFs Sarcolambin A and Sarcolambin B (28aa, 29aa, respectively) are orthologs of human Phospholamban and Sarcolipin[59], and were previously identified by functional conservation rather than sequence conservation. Future studies comparing fly-human smORFs using new structural comparison tools[60] may reveal additional conserved smORFs.

Similarly, our resource of reagents to modify smORF function in Drosophila is incomplete. For example, of the 298 conserved smORF genes, the TRiP/DRSC generated 165 sgRNA-KO lines (55%) and 176 sgRNA-OE lines (59%) (as of October 2023). Furthermore, it is unknown how well each sgRNA reagent works, which can be affected by gene expression[61], local chromatin[62], or genetic variation[63]. Finally, for our 25 homozygous KO smORF fly strains, we did not confirm gene KO by antibody staining, nor analyze neighboring gene expression.

STAR★Methods:

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Susan Celniker (secelniker@lbl.gov).

Materials availability—Fly stocks generated in this study have been deposited at the Bloomington Drosophila Stock Center, described in Supplemental File 7. sgRNA-KO and sgRNA-OE plasmids generated in this study are available by request from the DRSC/TRiP, described in Supplemental File 7.

Data and code availability

- Data availability: Raw sequencing data is available at the NCBI Short Read Archive (SRA): SRR18575339, SRR18575340, SRR18575342, SRR18575343, SRR18575345 and SRR18575346.
- Code availability: This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Flies were maintained on standard fly food at 25°C unless otherwise noted. Fly stocks were obtained from the Celniker lab collection, Perrimon lab collection, Bloomington Drosophila Stock center (indicated with BL#), or generated in this study (see below).

Celniker lab stocks:

Oregon-R

Perrimon Lab stocks:

yv; Gla/CyO

yw; nos-Cas9attP40/CyO

yw;; nos-Cas9attP2

lethal/FM7,GFP

yw; Gla/CyO

yw;; TM3, Sb/TM6b

lethal/FM7,GFP;; TM3, Ser

yw; Sp/CyO; MKRS/TM6B

Bloomington Stocks:

sgRNA lines (see Supplemental File 7)

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

Actin-Cas9 (BL54590)

attP40 (BL36304)

Mhc>Cas9 (BL67079)

LSP>Cas9 (BL67087)

Myo1a>Cas9 (BL67088)

Pnr>Cas9 (BL67077)

Nub>Cas9 (BL67086)

Elav>Cas9 (BL67073)

yw (BL1495)

yv nanos-phiC31; attP40 (BL25709)

y vas-phiC31; attP VK00037 (BL24872)

y vas-phiC31;; attP VK00033 (BL24871)

Information on the smORF KO and UAS-cDNA stocks are described in Supplemental File 7, including Bloomington Stock #s.

Age and sex of flies used in this study:

Figure 3 – Mixed males and female embryos imaged at stages 13–16.

Figure 4 – Mixed males and female embryos collected at 0–24 hr.

Figures 5,6; Figure S4A–F – Mixed males and female adults scored 0–7 days after eclosion (~10days-17days old)

Figure S4G; Figure S5 - Mixed males and female adults scored 7 days after eclosion (17days old)

METHOD DETAILS

Bioinformatic identification of 298 fly-human conserved smORFs—First, 266,066 human smORFs were selected, including all annotated human smORF transcripts (18,494 annotated in GENCODE version 24 less than 100aa), and 215,901 smORFs identified by 3-frame translations of all human transcripts that lack298 long ORFs (11–100aa). This set was filtered to identify high-confidence elements by leveraging a stringent, high-confidence set of conserved *Drosophila* smORFs. In *Drosophila*, there are 960 genes

encoding unique peptides with no more than 100 aa from protein-coding genes annotated at FlyBase (release 6.49) with evidence of translation. This set was expanded by taking smORFs (11–100aa) predicted from two independent studies[66], adding 2,819 additional smORFs with evidence of translation from either Ribosome Profiling or conservation among Drosophilidae.

For our ortholog discovery workflow, see Fig. 1A. Specifically, we used Diopt v8 and 9 (https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl)[67] for ortholog analysis. In parallel and to corroborate results, we also used deltablast 2.9.0+ build Sep 30 2019 01:57:31 with the following parameters: (Matrix : BLOSUM62); (Gap Penalties: Existence: 11, Extension :1); (Neighboring words threshold: 11); (Window for multiple hits:40). We then filtered the deltablast results for *D. rerio*, *C. elegans*, *A. thaliana* proteins using the following cutoffs <=250aa and E-value <=10^{-1}. The *D. rerio*, *C. elegans* and *A. thaliana* peptide sequence files used, are as follow: Danio_rerio.GRCz11.pep.all.fa, Caenorhabditis_elegans.WBcel235.pep.all.fa (https://ftp.ensemblgenomes.ebi.ac.uk/pub/metazoa/release-56/fasta/caenorhabditis_elegans/) and TAIR10_pep_20101214.faa (https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FProteins%2FTAIR10_protein_lists). These analyses identified orthologs in human for 291 fly genes, and an additional 7 were discovered using sim3 analysis[68].

Clustering of ontological anatomical annotations of embryonic expression patterns—Embryos were clustered using a bagged and cross validated procedure to ensure cluster stability. The importance of stability was made apparent to us when, during the course of our study, one additional smORF was discovered and added to this analysis, and the resulting clusters differed significantly. We stabilized the clustering as follows:

First, we subsampled 80% of smORFs and used hierarchical clustering with Ward linkage to form candidate clusters, holding out 20%. Here, the hold-out is to assess stability by inducing some randomness -- note that this differs from the use of a holdout in supervised learning. We selected 10K 80/20 splits. Cluster number was selected using the Gap Statistic -- the first local maximum value was selected. We computed a proximity matrix for consensus clustering as follows: for each pair of smORFs we recorded the fraction of our 10K bagged clusterings in which they appeared in the same cluster. This matrix was then input to hierarchical clustering with Ward linkage, and the Gap Statistic was used to select cluster number as above.

Embryo Collections for Coordinated Ribosome and RNA-Seq Profiling—Embryos from ~14 g of Oregon-R flies were collected on standard molasses collection trays after flies were acclimated to the environmental conditions of the cage (2°C and 70% humidity) for three days. Six, two-hour embryonic time periods (0–2, 2–4 hr., 4–6 hr., 10–12 hr., 14–16 hr., and 16–18 hr. were collected simultaneously, which allowed for immediate RNA-seq library and ribosome profiling construction of all six stages. See Supplemental File 6 for step-by-step protocol.

RNA Preparation and Sequencing Methods—Embryos were homogenized using a Pellet Pestle Cordless Motor (Kimble Cat. No. 749540–0000; Pellet Pestles Sigma Cat. No. Z359947), RNA was extracted using TRIzol Reagent (Thermo Fisher, Cat. No. 15596026) and purified with the RNeasy Mini Kit (QIAGEN Cat. No. 74106). Libraries were constructed with the NEBNext Ultra Directional RNA Library Prep Kit for Illumina according to manufacturer's recommendations (NEB, cat. no. E7420) using 14 cycles of PCR. Libraries were sequenced on the Illumina NovaSeq 6000. Raw sequencing data is available at the NCBI Short Read Archive (SRA): SRR18575339, SRR18575340, SRR18575342, SRR18575343, and SRR18575345 and SRR18575346.

We used the STAR aligner v2.73a to align RNA-seq data to the *D. melanogaster* genome (Rel 6). The picard v2.20.1 MarkDuplicates tool was used to remove PCR duplicates and the deduplicated BAM alignment files were converted to bigWig format using a custom tool and the UCSC bedGraphToBigWig tool. We ran fastp 0.20.1 to get FASTQ file statistics.

Ribosome Profiling Methods—Polysome profiling was performed on all six samples. Briefly, embryos from each time period (i.e., sample) were treated with harringtonine (LKT Laboratories, H0169) in mild lysis buffer followed by the addition of cycloheximide and immediate grinding. Samples were subjected to a 10%–50% sucrose gradient and all polysomes were collected and combined. Collected polysome fractions were pelleted via sucrose cushion (34%) and subjected to RNaseI digestion. After resuspension another cushion (34%) was performed and RNA was coprecipitated with GlycoBlue (Thermo Fisher, cat. No. AM95150). Recovered RNA was then run on a 15% TBE-Urea Gel (Thermo Fisher cat. no. EC68852BOX), followed by gel size selection to isolate ribosome protected fragments (26–31 nt) (ZR small-RNA PAGE Recovery Kit, Zymo Research, cat. no. R1070). Following end-repair phosphorylation of RNA molecules, libraries were constructed with the NEBNext Small RNA Library Preparation Kit according to manufacturer's recommendations (NEB, cat. no. E7330). See Supplemental File 6 for step-by-step protocol.

Sequencing, read processing and mapping—Ribo-seq libraries were sequenced with the Illumina NovaSeq 6000. Raw sequencing data is available at the SRA: SRR18575338, SRR18575341, SRR18575344, SRR18575347, SRR18575348, and SRR18575353. Read processing and mapping were performed on an Ubuntu 18.04 Linux cluster running Kubernetes v1.16. on 240 total cores with 1500 GB of total RAM. Reads were processed with the following commands:

```
QC of raw reads with FastQC (http://www.bioinformatics.babraham.ac.uk/
projects/fastqc/)
       fastqc file_one.fastq.gz file_two.fastq.gz
Clip adapters from Illumina raw reads with the following commands with
Cutadapt[69]
      cutadapt -a adapter sequence for file_one.fastq.gz \
      -A adapter sequence for file_one.fastq.gz -j 10  \
```

```
      -o file_one_clipped.fastq -p file_two_clipped.fastq  \
      file_one.fastq.gz file_two.fastq.gz
```

Trim clipped reads based on position quality with the following commands:

```
      cutadapt -q 33 -j 12 -o file_one_clipped_trimmed.fastq
file_one_clipped.fastq
      cutadapt -q 33 -j 12 -o file_two_clipped_trimmed.fastq
file_two_clipped.fastq
```

QC of processed reads:

```
      fastqc file_one_clipped_trimmed.fastq  \ file_two_clipped_trimmed.fastq
```

Remove reads smaller than 26 bp and larger than 31 bp:

```
      cutadapt --pair-filter=any --minimum-length=26 --maximum-length=31 -j
20  \
      -o file_one_clipped_trimmed_min_max_removed.fastq  \
      -p file_two_clipped_trimmed_min_max_removed.fastq  \
file_one_clipped_trimmed.fastq file_two_clipped_trimmed.fastq
```

Map first to the rDNA reference with Bowtie2[70] to 1) remove rRNA sequences and decrease mapping time to nuclear reference genome, and 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
```

Map non-rRNA reads to nuclear reference genome with STAR[71]:

```
      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
```

Detection of smORFs with Ribosome Profiling—Identification of translated sequences was performed using ORFquant[72]. For each predicted ORF with mapped reads we recorded all ORFquant summary statistics. We classified ORFs as detected if the adjusted P-value was less than or equal to 0.05 (Supplemental Table 5).

Gene enrichment analyses—GO and KEGG enrichment analyses were performed with g:Profiler[64].

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, and Phusion polymerase (M-0530, NEB) was used when DNA fragments were sequenced or used for molecular cloning. DNA fragments were run on a 1% agarose gel for imaging or purified on QIAquick columns (28115, Qiagen) for sequencing analysis. Sanger sequencing was performed at the DF/HCC DNA Resource Core facility and chromatograms were analyzed using Lasergene 13 software (DNASTAR).

For isolating flies with frameshift indels in smORF genes, the target site was PCR amplified from single fly genomic DNA and PCR fragments were Sanger sequenced. For isolating flies with whole gene deletion of smORF genes by dual sgRNA cutting, the target region was PCR amplified from single fly genomic DNA. Primers were designed to flank the two sgRNA cut sites, such that a deletion of the intervening sequence would produce a clear band size difference on an agarose gel. Deletion PCR fragments were Sanger sequenced. Genotyping primer sequences are listed in Supplemental File 7.

For RT-qPCR analysis of smORF overexpression by CRISPRa, adult flies (tub-Gal4, UAS-dCas9-VPR, sgRNA-OE) were flash frozen in liquid nitrogen. 4–14 frozen flies (equal mixture of males and females) were homogenized in 1000ul Trizol (Invitrogen 15596026), RNA partially purified by chloroform extraction, and RNA extracted using a Direct-zol RNA Miniprep kit (Zymo Research, R2050). cDNA was generated using the iScript Reverse Transcription Supermix (BioRad 1708840). cDNA was analyzed by RT-qPCR using iQ SYBR Green Supermix (BioRad 170–8880) on a CFX96 Real-Time system (BioRad). qPCR primer sequences are listed in Supplemental File 7. Each qPCR reaction was performed with five biological replicates (except *CG14818*, which had two biological replicates), with two technical replicates each. Data from smORF specific primers were normalized to primers that amplify GAPDH. Statistical significance was calculated using a T-Test.

Molecular cloning—Plasmid DNAs were constructed and propagated using standard protocols. Briefly, chemically competent TOP10 E.coli. (Invitrogen, C404010) were transformed with plasmids containing either Ampicillin or Kanamycin resistance genes and were selected on LB-Agar plates with 100µg/ml Ampicillin or 50µg/ml Kanamycin. Oligo sequences are in Supplemental File 7.

sgRNA expression plasmids: Plasmids encoding sgRNAs were generated using previously described protocols. sgRNAs were designed using the Find CRISPR tool (https://

www.flyrnai.org/crispr3/web) for optimal predicted cutting activity[42]. For sgRNAs cloned into *pCFD3*[73], annealed oligos encoding a sgRNA spacer were ligated (T4 DNA ligase, NEB, M0202S) into *pCFD3* digested with BbsI (NEB, R3539). For sgRNAs cloned into *pCFD4*[73], dual sgRNAs were PCR amplified from *pCFD4* template, and inserted by Gibson assembly (NEB, E2611) into *pCFD4* digested with BbsI. For sgRNAs cloned into *pCFD5*[45], dual sgRNAs were PCR amplified from *pCFD5* template, and inserted by Gibson assembly into *pCFD5* digested with BbsI. sgRNAs cloned into *pCFD3* and were performed by DRSC/TRiP (https://fgr.hms.harvard.edu/). Information on sgRNA-KO and sgRNA-OE plasmids generated by the DRSC/TRiP is available at https://www.flyrnai.org/tools/grna_tracker/web/. Information on remaining *pCFD4* and *pCFD5* sgRNA plasmids is in Supplemental File 7.

UAS-cDNA plasmids: UAS-cDNA plasmids were constructed as previously described[74,75]. Briefly, Entry plasmids used to clone into pGW-HA.attB[76] were generated by PCR amplifying coding sequence and inserting into pDONR223[77]. cDNA sequence was PCR amplified from BDGP cDNA gold clones[78]. Entry plasmids used to clone into *pWalium10-roe*[75] were generated by PCR amplifying coding sequence and inserting into *pEntr* using either dTopo (Invitrogen, K240020) or Gibson assembly (NEB, E2611). cDNA sequence was PCR amplified from cDNA reverse transcribed from total RNA (either S2R+ cell or adult fly) or adult fly genomic DNA. Entry clones were recombined into destination vectors (pGW-HA.attB[76] or *pWalium10-roe*[75]) using using LR Clonase II Enzyme mix (Invitrogen 11791–020).

Fly Genetics

Transgenic flies: Transgenic flies were generated by phiC31 integration of attB-containing plasmids into attP landing sites. sgRNA-expressing plasmids and pWalium10-cDNA plasmids were integrated into *attP40*, and pGW-HA-cDNA plasmids were integrated into *VK00037* or *VK00033*. Briefly, plasmid DNA was purified twice on QIAquick columns and eluted in injection buffer (100 μM NaPO4, 5 mM KCl) at a concentration of 200 ng/μL. Plasmid DNA was injected into ~50 fertilized embryos (e.g. yv nos-phiC31int; attP40) and resulting progeny were outcrossed to screen for transgenic founder progeny by scoring for *white+*.

smORF knockout by frameshift indel by transgenic crossing: Flies expressing a sgRNA that targets the 5’ coding sequence (see Supplemental File 7) were crossed with *nos-Cas9* flies. *nos-Cas9attP2* was used for targeting genes on Chromosomes X and II, and *nos-Cas9attP40* was used for targeting genes on chromosome III. F1 progeny were crossed with a balancer strain. Single fly F2 progeny were crossed with a balancer strain, taken for genotyping, and F2 crosses with a frameshift indel were kept and balanced and homozygosed if possible. Frameshift knockout lines were generated either by WellGenetics, Shu Kondo, or in the Perrimon lab.

smORF knockout by full gene deletion by injection: Plasmids encoding two sgRNAs that flank a gene locus were injected into *nos-Cas9* embryos. Injected F0 adults were crossed to with a balancer strain. Single fly F1 progeny were crossed with a balancer strain,

taken for genotyping, and F1 crosses with a full gene deletion were kept and balanced and homozygosed if possible.

smORF knockout by CRISPaint insertion by injection: To generate the bc10 KO allele, a plasmid encoding a sgRNA that targets the 5' coding sequence of bc10 (GP01409) was co-injected with pCRISPaint-T2A-Gal4-3xP3-RFP[79] (Addgene #127556) into *nos-Cas9attP40* embryos. Injected F0 adults were outcrossed to *yw*, single RFP+ F1 progeny were crossed with *yw;; TM3, Sb/TM6b,* and the insertion in bc10 was verified by PCR and sanger sequencing. The *bc10-CRISPaint* allele contains *T2A-Gal4* inserted in the reverse orientation relative to the 5'-3' *bc10* transcript, and thus is not a Gal4 reporter allele.

To generate double smORF knockout lines, smORF alleles on different chromosomes were brought into the same strain by outcrossing to double balancer lines.

CRISPR-KO F1 crosses and phenotyping: Of the 165 smORF genes targeted with at least sgRNAs for knockout, 11 genes were tested with two independent sgRNAs, and four genes were tested with three sgRNAs. Lines expressing sgRNAs that target smORF 5' coding sequence were crossed with a line ubiquitously expressing Cas9, *Act5c-Cas9*. Specifically, male sgRNA flies were used that were heterozygous with a balancer chromosome (CyO or TM3, Sb), and were crossed with homozygous *Act5c-Cas9* female flies. To quantify the viability of F1 flies with somatic KO, we recorded the number of balancer progeny (*Act5c-Cas9/Bal*) and non-balancer progeny (*sgRNA/Act5c-Cas9*). The total number of F1 progeny counted per cross was 920>n>31. To calculate a viability score, the number non-balancer flies was divided by the mendelian expected number of non-balancer flies (# F1 progeny/2) and multiplied by 100 (# observed non-balancer/# expected non-balancer*100). Negative control crosses were *attp40/CyO* males crossed with *Act5c-Cas9* females. For crosses leading to reduced viability, a Chi-square test was used to determine significance by comparing the # expected non-balancer flies from negative control (*attp40*) vs experimental (sgRNA) crosses. Those sgRNA hits that had significant Chi-square p-values (<0.001) with low viability (<25%) (Figure 5B, Supplemental File 7) were crossed with tissue specific Cas9 lines (Figure S4). Viability scores and Chi-square tests were performed similarly to *Act5c-Cas9* crosses.

CRISPRa F1 crosses and phenotyping: Of the 176 smORF genes targeted with at least sgRNAs for overexpression, 19 genes were tested with two independent sgRNAs, and one gene was tested with three sgRNAs. Lines expressing sgRNAs that target upstream of a smORF transcriptional start site (TSS) were crossed with line ubiquitously expressing dCas9-VPR, *tub>VPR* (*tub-Gal4, UAS-dCas9-VPR/S-T*). "S-T" are a second and third chromosome balancer pair that segregate together due to a reciprocal translocation, and are marked by Cy, Hu, and Tb (T(2;3)TSTL14, SM5: TM6B, Tb[1]). Male sgRNA flies were crossed with tub>VPR females. A viability score was calculated similar to CRISPR-KO F1 crosses, (# observed non-balancer/# expected non-balancer*100). When using homozygous sgRNA males, the expected number of non-balancer flies was # F1 progeny/2. When using heterozygous sgRNA/Bal males, the expected number of non-balancer flies was # F1 progeny/4. The total number of F1 progeny counted per cross was 346>n>56. Negative control crosses were *attp40/CyO* males crossed with *tub>VPR* females. Chi-square analysis

was performed similarly to CRISPR-KO F1 crosses. No crosses had significantly reduced viability.

cDNA overexpression crosses and phenotyping: Transgenic UAS-cDNA lines were crossed with a ubiquitous Gal4 line. For convenience, we used the same driver used for CRISPRa crosses, *tub>VPR* (*tub-Gal4, UAS-dCas9-VPR/S-T*). A viability score was calculated similar to CRISPRa F1 crosses. Negative control crosses were *attp40/CyO* males crossed with *tub>VPR* females. The total number of F1 progeny counted per cross was 706>n>100. Chi-square analysis was performed similarly to CRISPR-KO F1 crosses.

Stressful food recipes: For all food types, standard lab fly food was melted in a microwave, and distilled water (dH20) was added as 10% boiled volume (100ml boiled food + 10ml dH20) to replace the evaporated water. This is used as control food. For high salt food, solid NaCl (Fisher Scientific, S271) was added at 30% weight per volume of control food (e.g. 100ml control food + 30g NaCl) and mixed well. For high fat food, solid coconut oil (Sigma, W530155) was added at 30% weight per volume of control food (e.g. 100ml control food + 30g coconut oil) and mixed well. For starvation food, control food was diluted to 30% in 1% melted agar (BD, 214030) in 1x PBS (Gibco, 10010–023) (e.g. 30ml control food + 70ml melted 1% agar in 1x PBS). Melted liquid food types were poured into empty vials and cooled at 4°C.

Quantification of developmental timing on stressful food: 24hr larvae from homozygous viable smORF lines were transferred to stressful food or control food and raised on this food until pupal eclosion as adults. To increase the fecundity of the adult flies, four days prior to the 24hr larval collection, ~150 adult flies from each KO line were transferred to fresh food containing yeast paste. Adult flies were transferred onto fresh food bottles containing yeast paste and allowed to lay for 4hr. 24hr after the end of egg deposition, 30 freshly hatched larvae (24hr-28hr old) were transferred into vials containing either control food or stressful food (control, high fat, high salt, starvation). The time of pupariation and fly eclosion was determined once at least 15 flies pupariated and eclosed, respectively. yw flies were included as a negative control genotype. Each genotype-foodtype experiment was carried out in at least triplicate. For those genotype-foodtypes with a developmental delay, significance was calculated using a One-Way ANOVA test run with a Dunnet post-hoc test using GraphPad Prism.

Adult wing mounting, imaging, analysis: sgRNA-KO lines were crossed with *nub>Cas9* and the wings of adult progeny were removed using forceps under a dissecting microscope. For each genotype, at least 6 wings were collected. Removed wings were placed onto a drop of mounting medium (50% Permount (Fisher Scientific, SP15), 50% Xylenes (Fisher Scientific X5)) on a microscope slide (Thermo Scientific, 3050) and mounted using a coverslip (VWR, 48393059). The coverslip was sealed to the microscope slide with clear nail polish. Images of the wings were taken using a stereo microscope (Zeiss Axio Zoom V16) at 32x magnification.

Bioinformatics and literature searching—For protein alignments in Figure S6, we downloaded protein sequence files from Flybase.org or NCBI, aligned them using Clustal

Omega (https://www.ebi.ac.uk/Tools/msa/clustalo), and generated alignment images using JalView (https://www.jalview.org/).

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

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed using GraphPad Prism software (v9). P values represented in all Figures are P=0.05(*), P=0.01(**), P=0.001(***), P ≤ 0.0001 (****).

Quantification of fly viability from F1 crosses (sgRNA-KO, sgRNA-OE, UAS-cDNA) was performed by comparing the number of balancer progeny (e.g. *Act5c-Cas9/Bal*) and non-balancer progeny (e.g. *sgRNA/Act5c-Cas9*) (See Figures 5 B,D,F). The number of F1 progeny counted per cross was 918>n>33. Significance of fly viability crosses were performed using Chi-squared analysis, by comparing to negative control crosses involving the *attP40* transgenic landing site (e.g. *attP40* × *Act5c-Cas9*).

Quantification of developmental delay of homozygous KO fly lines on different food types was performed by recording the time of pupariation and fly eclosion of at least 15 flies, comparing to negative control *yw* flies (see Figure 6). Each genotype-foodtype experiment was carried out in at least triplicate. Significance was calculated using a One-Way ANOVA test run with a Dunnet post-hoc test. Error bars indicate SD.

Quantification of qPCR data was performed by comparing transcript expression to *Rp49* and *Gapdh* (see Figure S5). Statistical significance was calculated using a T-Test. For each genotype, N=5 biological replicates, except CG14818 sgRNA-OE which had 2 biological replicates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We thank Benjamin W Booth for processing the RNA-seq data and submitting RNA and ribosome profiling data to the SRA. We thank Marcus Stoiber for his initial identification of conserved smORFs. We thank Nick Ingolia and Jonathan Weissman for advice about ribosome profiling. We thank the Transgenic RNAi Project at Harvard Medical School (R24OD030002) for sgRNA-expressing fly lines. BioRender was used to prepare the graphical abstract. This work was funded by an award from the NIH National Human Genome Research Institute (R01HG009352) to S.E.C. (Principal Investigator). J.A.B. was supported by the Damon Runyon Foundation (DRG-2258-16) and a "Training Grant in Genetics" T32 Ruth Kirschstein-National Research Service Award institutional research training grant funded through the NIH/National Institute of General Medical Sciences (T32GM007748). NP is an HHMI investigator.

References

1. Plaza S, Menschaert G, and Payre F (2017). In Search of Lost Small Peptides. Annu Rev Cell Dev Biol 33, 391–416. 10.1146/annurev-cellbio-100616-060516. [PubMed: 28759257]
2. Couso JP, and Patraquim P (2017). Classification and function of small open reading frames. Nat Rev Mol Cell Biol 18, 575–589. 10.1038/nrm.2017.58. [PubMed: 28698598]
3. Guerra-Almeida D, Tschoeke DA, and Nunes-da-Fonseca R (2021). Understanding small ORF diversity through a comprehensive transcription feature classification. DNA Res 28. 10.1093/dnares/dsab007.
4. Jain N, Richter F, Adzhubei I, Sharp AJ, and Gelb BD (2023). Small open reading frames: a comparative genetics approach to validation. BMC Genomics 24, 226. 10.1186/s12864-023-09311-7. [PubMed: 37127568]
5. Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, and Saghatelian A (2020). Accurate annotation of human protein-coding small open reading frames. Nat Chem Biol 16, 458–468. 10.1038/s41589-019-0425-0. [PubMed: 31819274]
6. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. (2000). The genome sequence of Drosophila melanogaster. Science 287, 2185–2195. 10.1126/science.287.5461.2185. [PubMed: 10731132]
7. Pearson RK, Anderson B, and Dixon JE (1993). Molecular biology of the peptide hormone families. Endocrinol Metab Clin North Am 22, 753–774. [PubMed: 7907289]
8. Snyder SH, and Innis RB (1979). Peptide neurotransmitters. Annu Rev Biochem 48, 755–782. 10.1146/annurev.bi.48.070179.003543. [PubMed: 38738]
9. Yang P, Maguire JJ, and Davenport AP (2015). Apelin, Elabela/Toddler, and biased agonists as novel therapeutic agents in the cardiovascular system. Trends Pharmacol Sci 36, 560–567. 10.1016/j.tips.2015.06.002. [PubMed: 26143239]
10. Henras A, Henry Y, Bousquet-Antonelli C, Noaillac-Depeyre J, Gelugne JP, and Caizergues-Ferrer M (1998). Nhp2p and Nop10p are essential for the function of H/ACA snoRNPs. EMBO J 17, 7078–7090. 10.1093/emboj/17.23.7078. [PubMed: 9843512]
11. Bosch JA, Ugur B, Pichardo-Casas I, Rabasco J, Escobedo F, Zuo Z, Brown B, Celniker S, Sinclair DA, Bellen HJ, and Perrimon N (2022). Two neuronal peptides encoded from a single transcript regulate mitochondrial complex III in Drosophila. Elife 11. 10.7554/eLife.82709.
12. Rathore A, Chu Q, Tan D, Martinez TF, Donaldson CJ, Diedrich JK, Yates JR 3rd, and Saghatelian A (2018). MIEF1 Microprotein Regulates Mitochondrial Translation. Biochemistry 57, 5564–5575. 10.1021/acs.biochem.8b00726. [PubMed: 30215512]
13. Stein CS, Jadiya P, Zhang X, McLendon JM, Abouassaly GM, Witmer NH, Anderson EJ, Elrod JW, and Boudreau RL (2018). Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. Cell Rep 23, 3710–3720 e3718. 10.1016/j.celrep.2018.06.002. [PubMed: 29949756]
14. Savard J, Marques-Souza H, Aranda M, and Tautz D (2006). A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. Cell 126, 559–569. 10.1016/j.cell.2006.05.053. [PubMed: 16901788]
15. Galindo MI, Pueyo JI, Fouix S, Bishop SA, and Couso JP (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol 5, e106. 10.1371/journal.pbio.0050106. [PubMed: 17439302]
16. Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, and Kageyama Y (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science 329, 336–339. 10.1126/science.1188158. [PubMed: 20647469]
17. Chanut-Delalande H, Hashimoto Y, Pelissier-Monier A, Spokony R, Dib A, Kondo T, Bohere J, Niimi K, Latapie Y, Inagaki S, et al. (2014). Pri peptides are mediators of ecdysone for the temporal control of development. Nat Cell Biol 16, 1035–1044. 10.1038/ncb3052. [PubMed: 25344753]

18. Chanet S, and Schweisguth F (2012). Regulation of epithelial polarity by the E3 ubiquitin ligase Neuralized and the Bearded inhibitors in Drosophila. Nat Cell Biol 14, 467–476. 10.1038/ncb2481. [PubMed: 22504274]
19. Lai EC, Bodner R, Kavaler J, Freschi G, and Posakony JW (2000). Antagonism of notch signaling activity by members of a novel protein family encoded by the bearded and enhancer of split gene complexes. Development 127, 291–306. 10.1242/dev.127.2.291. [PubMed: 10603347]
20. Bardin AJ, and Schweisguth F (2006). Bearded family members inhibit Neuralized-mediated endocytosis and signaling activity of Delta in Drosophila. Dev Cell 10, 245–255. 10.1016/j.devcel.2005.12.017. [PubMed: 16459303]
21. Hanyu-Nakamura K, Sonobe-Nojima H, Tanigawa A, Lasko P, and Nakamura A (2008). Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. Nature 451, 730–733. 10.1038/nature06498. [PubMed: 18200011]
22. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, and Obermayer B (2015). Extensive identification and analysis of conserved small ORFs in animals. Genome Biol 16, 179. 10.1186/s13059-015-0742-x. [PubMed: 26364619]
23. Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, and Couso JP (2011). Hundreds of putatively functional small open reading frames in Drosophila. Genome Biol 12, R118. 10.1186/gb-2011-12-11-r118. [PubMed: 22118156]
24. Fabre B, Choteau SA, Duboe C, Pichereaux C, Montigny A, Korona D, Deery MJ, Camus M, Brun C, Burlet-Schiltz O, et al. (2022). In Depth Exploration of the Alternative Proteome of Drosophila melanogaster. Front Cell Dev Biol 10, 901351. 10.3389/fcell.2022.901351. [PubMed: 35721519]
25. Patraquim P, Magny EG, Pueyo JI, Platero AI, and Couso JP (2022). Translation and natural selection of micropeptides from long non-canonical RNAs. Nat Commun 13, 6515. 10.1038/s41467-022-34094-y. [PubMed: 36316320]
26. Zhang M, Zhao J, Li C, Ge F, Wu J, Jiang B, Song J, and Song X (2022). csORF-finder: an effective ensemble learning framework for accurate identification of multi-species coding short open reading frames. Brief Bioinform 23. 10.1093/bib/bbac392.
27. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, and Tanabe M (2021). KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 49, D545–D551. 10.1093/nar/gkaa970. [PubMed: 33125081]
28. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. (2014). Diversity and dynamics of the Drosophila transcriptome. Nature 512, 393–399. 10.1038/nature12962. [PubMed: 24670639]
29. Rawlings ND, Tolle DP, and Barrett AJ (2004). Evolutionary families of peptidase inhibitors. Biochem J 378, 705–716. 10.1042/BJ20031825. [PubMed: 14705960]
30. Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, and Nielsen H (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 40, 1023–1025. 10.1038/s41587-021-01156-3. [PubMed: 34980915]
31. Brown CJ, Kaufman T, Trinidad JC, and Clemmer DE (2018). Proteome changes in the aging Drosophila melanogaster head. Int J Mass Spectrom 425, 36–46. 10.1016/j.ijms.2018.01.003. [PubMed: 30906200]
32. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. (2011). The developmental transcriptome of Drosophila melanogaster. Nature 471, 473–479. 10.1038/nature09715. [PubMed: 21179090]
33. Patraquim P, Mumtaz MAS, Pueyo JI, Aspden JL, and Couso JP (2020). Developmental regulation of canonical and small ORF translation from mRNAs. Genome Biol 21, 128. 10.1186/s13059-020-02011-5. [PubMed: 32471506]
34. Li H, Hu C, Bai L, Li H, Li M, Zhao X, Czajkowsky DM, and Shao Z (2016). Ultra-deep sequencing of ribosome-associated poly-adenylated RNA in early Drosophila embryos reveals hundreds of conserved translated sORFs. DNA Res 23, 571–580. 10.1093/dnares/dsw040. [PubMed: 27559081]

35. Chen CL, Hu Y, Udeshi ND, Lau TY, Wirtz-Peitz F, He L, Ting AY, Carr SA, and Perrimon N (2015). Proteomic mapping in live Drosophila tissues using an engineered ascorbate peroxidase. Proc Natl Acad Sci U S A 112, 12093–12098. 10.1073/pnas.1515623112. [PubMed: 26362788]

36. Aradska J, Bulat T, Sialana FJ, Birner-Gruenberger R, Erich B, and Lubec G (2015). Gel-free mass spectrometry analysis of Drosophila melanogaster heads. Proteomics 15, 3356–3360. 10.1002/pmic.201500092. [PubMed: 26201256]

37. Cammarato A, Ahrens CH, Alayari NN, Qeli E, Rucker J, Reedy MC, Zmasek CM, Gucek M, Cole RN, Van Eyk JE, et al. (2011). A mighty small heart: the cardiac proteome of adult Drosophila melanogaster. PLoS One 6, e18497. 10.1371/journal.pone.0018497. [PubMed: 21541028]

38. Casas-Vila N, Bluhm A, Sayols S, Dinges N, Dejung M, Altenhein T, Kappei D, Altenhein B, Roignant JY, and Butter F (2017). The developmental proteome of Drosophila melanogaster. Genome Res 27, 1273–1285. 10.1101/gr.213694.116. [PubMed: 28381612]

39. Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, and Karr TL (2006). Genomic and functional evolution of the Drosophila melanogaster sperm proteome. Nat Genet 38, 1440–1445. 10.1038/ng1915. [PubMed: 17099714]

40. Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, Polpitiya A, Petritis K, and Karr TL (2010). The Drosophila melanogaster sperm proteome-II (DmSP-II). J Proteomics 73, 2171–2185. 10.1016/j.jprot.2010.09.002. [PubMed: 20833280]

41. Port F, Strein C, Stricker M, Rauscher B, Heigwer F, Zhou J, Beyersdorffer C, Frei J, Hess A, Kern K, et al. (2020). A large-scale resource for tissue-specific CRISPR mutagenesis in Drosophila. Elife 9. 10.7554/eLife.53865.

42. Zirin J, Hu Y, Liu L, Yang-Zhou D, Colbeth R, Yan D, Ewen-Campen B, Tao R, Vogt E, VanNest S, et al. (2020). Large-Scale Transgenic Drosophila Resource Collections for Loss- and Gain-of-Function Studies. Genetics 214, 755–767. 10.1534/genetics.119.302964. [PubMed: 32071193]

43. Ewen-Campen B, Yang-Zhou D, Fernandes VR, Gonzalez DP, Liu LP, Tao R, Ren X, Sun J, Hu Y, Zirin J, et al. (2017). Optimized strategy for in vivo Cas9-activation in Drosophila. Proc Natl Acad Sci U S A 114, 9409–9414. 10.1073/pnas.1707635114. [PubMed: 28808002]

44. Jia Y, Xu RG, Ren X, Ewen-Campen B, Rajakumar R, Zirin J, Yang-Zhou D, Zhu R, Wang F, Mao D, et al. (2018). Next-generation CRISPR/Cas9 transcriptional activation in Drosophila using flySAM. Proc Natl Acad Sci U S A 115, 4719–4724. 10.1073/pnas.1800677115. [PubMed: 29666231]

45. Port F, and Bullock SL (2016). Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs. Nat Methods 13, 852–854. 10.1038/nmeth.3972. [PubMed: 27595403]

46. Zirin J, Nieuwenhuis J, Samsonova A, Tao R, and Perrimon N (2015). Regulators of autophagosome formation in Drosophila muscles. PLoS Genet 11, e1005006. 10.1371/journal.pgen.1005006. [PubMed: 25692684]

47. Birse RT, Choi J, Reardon K, Rodriguez J, Graham S, Diop S, Ocorr K, Bodmer R, and Oldham S (2010). High-fat-diet-induced obesity and heart dysfunction are regulated by the TOR pathway in Drosophila. Cell Metab 12, 533–544. 10.1016/j.cmet.2010.09.014. [PubMed: 21035763]

48. Stergiopoulos K, Cabrero P, Davies SA, and Dow JA (2009). Salty dog, an SLC5 symporter, modulates Drosophila response to salt stress. Physiol Genomics 37, 1–11. 10.1152/physiolgenomics.90360.2008. [PubMed: 19018044]

49. Zhang S, Reljic B, Liang C, Kerouanton B, Francisco JC, Peh JH, Mary C, Jagannathan NS, Olexiouk V, Tang C, et al. (2020). Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. Nat Commun 11, 1312. 10.1038/s41467-020-14999-2. [PubMed: 32161263]

50. Fecher C, Trovo L, Muller SA, Snaidero N, Wettmarshausen J, Heink S, Ortiz O, Wagner I, Kuhn R, Hartmann J, et al. (2019). Cell-type-specific profiling of brain mitochondria reveals functional and molecular diversity. Nat Neurosci 22, 1731–1742. 10.1038/s41593-019-0479-z. [PubMed: 31501572]

51. Wang J, Al-Ouran R, Hu Y, Kim SY, Wan YW, Wangler MF, Yamamoto S, Chao HT, Comjean A, Mohr SE, et al. (2017). MARRVEL: Integration of Human and Model Organism Genetic

Resources to Facilitate Functional Annotation of the Human Genome. Am J Hum Genet 100, 843–853. 10.1016/j.ajhg.2017.04.010. [PubMed: 28502612]

52. Bersuker K, Peterson CWH, To M, Sahl SJ, Savikhin V, Grossman EA, Nomura DK, and Olzmann JA (2018). A Proximity Labeling Strategy Provides Insights into the Composition and Dynamics of Lipid Droplet Proteomes. Dev Cell 44, 97–112 e117. 10.1016/j.devcel.2017.11.020. [PubMed: 29275994]
53. Remmelzwaal S, Geisler F, Stucchi R, van der Horst S, Pasolli M, Kroll JR, Jarosinska OD, Akhmanova A, Richardson CA, Altelaar M, et al. (2021). BBLN-1 is essential for intermediate filament organization and apical membrane morphology. Curr Biol 31, 2334–2346 e2339. 10.1016/j.cub.2021.03.069. [PubMed: 33857431]
54. Beall CJ, and Fyrberg E (1991). Muscle abnormalities in Drosophila melanogaster heldup mutants are caused by missing or aberrant troponin-I isoforms. J Cell Biol 114, 941–951. 10.1083/jcb.114.5.941. [PubMed: 1908472]
55. Henderson CA, Gomez CG, Novak SM, Mi-Mi L, and Gregorio CC (2017). Overview of the Muscle Cytoskeleton. Compr Physiol 7, 891–944. 10.1002/cphy.c160033. [PubMed: 28640448]
56. Lim SC, Smith KR, Stroud DA, Compton AG, Tucker EJ, Dasvarma A, Gandolfo LC, Marum JE, McKenzie M, Peters HL, et al. (2014). A founder mutation in PET100 causes isolated complex IV deficiency in Lebanese individuals with Leigh syndrome. Am J Hum Genet 94, 209–222. 10.1016/j.ajhg.2013.12.015. [PubMed: 24462369]
57. Dennerlein S, Poerschke S, Oeljeklaus S, Wang C, Richter-Dennerlein R, Sattmann J, Bauermeister D, Hanitsch E, Stoldt S, Langer T, et al. (2021). Defining the interactome of the human mitochondrial ribosome identifies SMIM4 and TMEM223 as respiratory chain assembly factors. Elife 10. 10.7554/eLife.68213.
58. Liang C, Zhang S, Robinson D, Ploeg MV, Wilson R, Nah J, Taylor D, Beh S, Lim R, Sun L, et al. (2022). Mitochondrial microproteins link metabolic cues to respiratory chain biogenesis. Cell Rep 40, 111204. 10.1016/j.celrep.2022.111204. [PubMed: 35977508]
59. Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, and Couso JP (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science 341, 1116–1120. 10.1126/science.1238802. [PubMed: 23970561]
60. Hamamsy T, Morton JT, Blackwell R, Berenberg D, Carriero N, Gligorijevic V, Strauss CEM, Leman JK, Cho K, and Bonneau R (2023). Protein remote homology detection and structural alignment using deep learning. Nat Biotechnol. 10.1038/s41587-023-01917-2.
61. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517, 583–588. 10.1038/nature14136. [PubMed: 25494202]
62. Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, Cheng AW, Trevino AE, Konermann S, Chen S, et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol 32, 670–676. 10.1038/nbt.2889. [PubMed: 24752079]
63. Housden BE, Valvezan AJ, Kelley C, Sopko R, Hu Y, Roesel C, Lin S, Buckner M, Tao R, Yilmazel B, et al. (2015). Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. Sci Signal 8, rs9. 10.1126/scisignal.aab3729. [PubMed: 26350902]
64. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, and Vilo J (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 47, W191–W198. 10.1093/nar/gkz369. [PubMed: 31066453]
65. Weiszmann R, Hammonds AS, and Celniker SE (2009). Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. Nat Protoc 4, 605–618. 10.1038/nprot.2009.55. [PubMed: 19360017]
66. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MA, Brocard M, and Couso JP (2014). Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife 3, e03528. 10.7554/eLife.03528. [PubMed: 25144939]

67. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, and Mohr SE (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12, 357. 10.1186/1471-2105-12-357. [PubMed: 21880147]
68. Chao KM, Zhang J, Ostell J, and Miller W (1997). A tool for aligning very similar DNA sequences. Comput Appl Biosci 13, 75–80. 10.1093/bioinformatics/13.1.75. [PubMed: 9088712]
69. Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 17, 3. 10.14806/ej.17.1.200.
70. Langmead B, and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. 10.1038/nmeth.1923. [PubMed: 22388286]
71. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. 10.1093/bioinformatics/bts635. [PubMed: 23104886]
72. Calviello L, Hirsekorn A, and Ohler U (2020). Quantification of translation uncovers the functions of the alternative transcriptome. Nat Struct Mol Biol 27, 717–725. 10.1038/s41594-020-0450-4. [PubMed: 32601440]
73. Port F, Chen HM, Lee T, and Bullock SL (2014). Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proc Natl Acad Sci U S A 111, E2967–2976. 10.1073/pnas.1405500111. [PubMed: 25002478]
74. Tang HW, Spirohn K, Hu Y, Hao T, Kovacs IA, Gao Y, Binari R, Yang-Zhou D, Wan KH, Bader JS, et al. (2023). Next-generation large-scale binary protein interaction network for Drosophila melanogaster. Nat Commun 14, 2162. 10.1038/s41467-023-37876-0. [PubMed: 37061542]
75. Perkins LA, Holderbaum L, Tao R, Hu Y, Sopko R, McCall K, Yang-Zhou D, Flockhart I, Binari R, Shim HS, et al. (2015). The Transgenic RNAi Project at Harvard Medical School: Resources and Validation. Genetics 201, 843–852. 10.1534/genetics.115.180208. [PubMed: 26320097]
76. Bischof J, Bjorklund M, Furger E, Schertel C, Taipale J, and Basler K (2013). A versatile platform for creating a comprehensive UAS-ORFeome library in Drosophila. Development 140, 2434–2442. 10.1242/dev.088757. [PubMed: 23637332]
77. Rual JF, Hirozane-Kishikawa T, Hao T, Bertin N, Li S, Dricot A, Li N, Rosenberg J, Lamesch P, Vidalain PO, et al. (2004). Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res 14, 2128–2135. 10.1101/gr.2973604. [PubMed: 15489335]
78. Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, et al. (2002). A Drosophila full-length cDNA resource. Genome Biol 3, RESEARCH0080. 10.1186/gb-2002-3-12-research0080. [PubMed: 12537569]
79. Bosch JA, Colbeth R, Zirin J, and Perrimon N (2020). Gene Knock-Ins in Drosophila Using Homology-Independent Insertion of Universal Donor Plasmids. Genetics 214, 75–89. 10.1534/genetics.119.302819. [PubMed: 31685521]

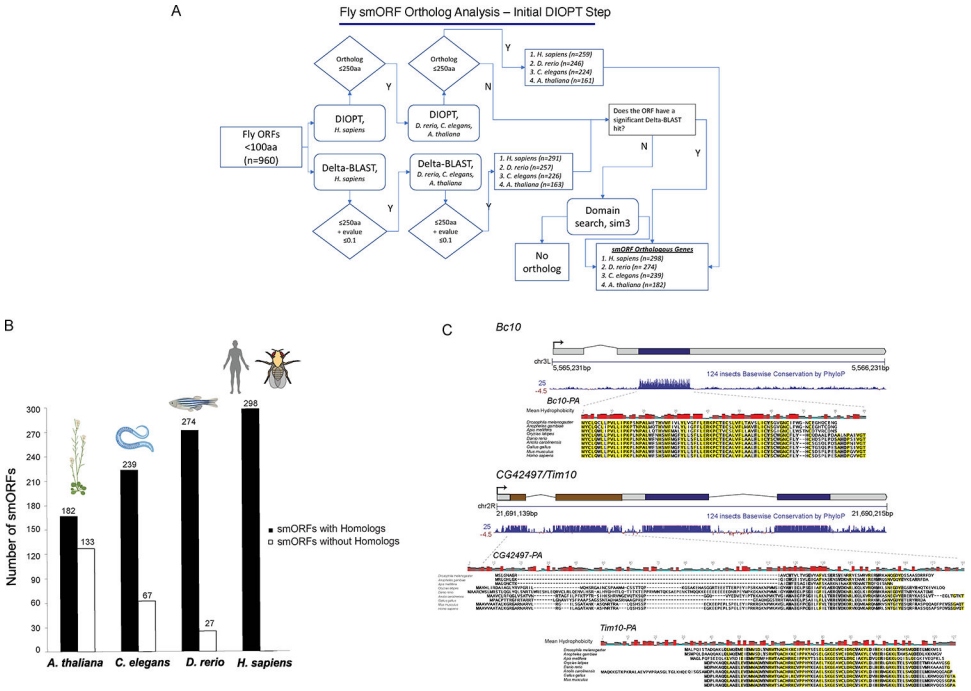


Figure 1. Conservation of smORF dataset.

(*A*) Flowthrough of bioinformatic identification of 298 fly-human conserved smORFs **(*B*)** Number of conserved smORFs in dataset with and without homologs in a selection of species with well-annotated transcriptomes. **(*C*)** Multiple-species alignments of conserved smORFs, including mean amino acid hydrophobicity at each alignment position. *bc10* (upper transcript) is encodes one smORF, whereas *CG42497* and *Tim10* (lower transcript) is polycistronic. See also Supplemental File 1 and Supplemental Figure S1.

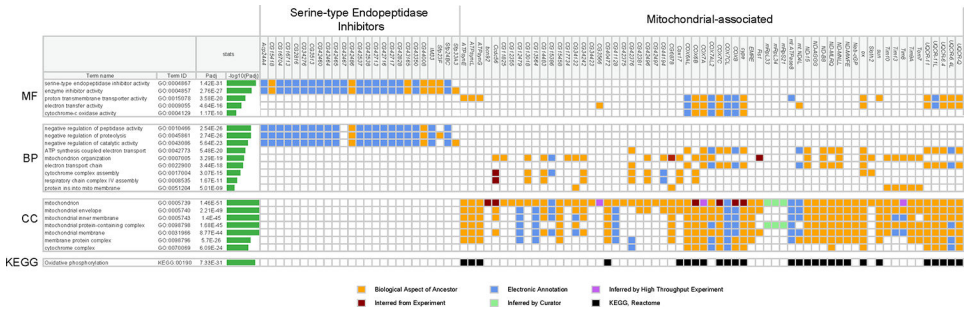


Figure 2. Gene Ontology (GO) and KEGG enrichment analysis of conserved smORFs.

Significantly enriched GO terms for molecular function (“MF”), biological process (“BP”), cellular component (“CC”) are plotted. GO and KEGG enrichment analyses were performed with g:Profiler[64]. Significantly enriched terms $<10^{-5}$ are shown that also encompass all conserved smORFs classified as serine-type endopeptidase inhibitors and mitochondria-associated conserved smORFs. See also Supplemental File 2.

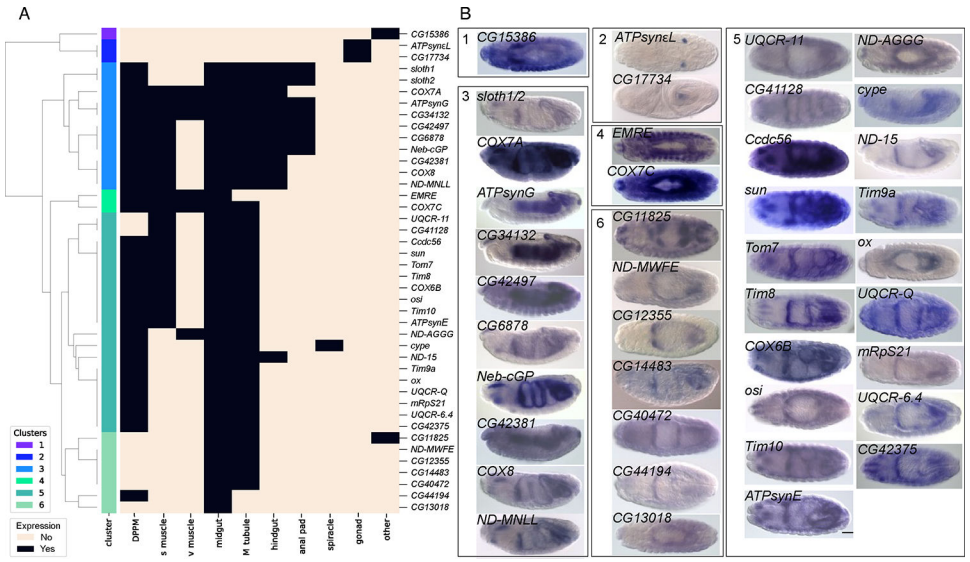


Figure 3. *In situ* mRNA hybridization patterns of mitochondria-associated smORFs.

(*A*) Clustering of mitochondria-associated conserved smORF *in situ* mRNA expression patterns[65]. For each mitochondrial conserved smORF, the organs where expression patterns were assigned are represented by red boxes, while blue boxes represent no annotated expression. Expression patterns across embryo stages are collapsed. “DPPM” = dorsal prothoracic pharyngeal muscle; “MT” = Malpighian tubules; “VNC” = ventral nerve cord; “D and V epidermis” = dorsal and ventral epidermis. **(*B*)** In-situ mRNA hybridization images for each mitochondria-associated conserved smORF with patterned expression. Each image was taken between embryonic stages 13–16. Scale bar is 50μm. See also Supplemental Files 3 and 4.

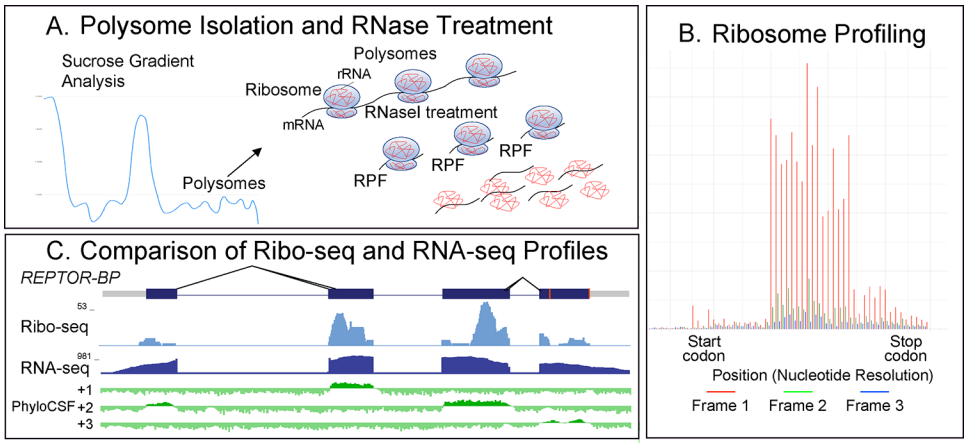


Figure 4. Ribosome Profiling.

(A) Overview of ribosome profiling workflow, where polysomes are isolated followed by digestion of inter-ribosome RNA. Ribosome protected fragments are then collected. **(*B*)** After sequencing, the number of in frame reads are analyzed to determine if RPFs were successfully sequenced. Distribution of tags per million (TPM) for six, embryonic time periods. **(*C*)** Comparison of ribosome profiling sequencing to mRNA-seq sequencing showing ribosome profiling libraries are constrained to CDS while mRNA libraries map to the entire annotated transcript. *REPTOR-BP* encodes four small peptides, 93, 94, 117 and 118aa (blue boxes). These peptides share the same translation start site (arrowhead) and differ by the addition of a glutamine (indicated by the different splice sites in the second exon) and by carboxy termini (indicated by alternate splice sites and red bars). See also Supplemental Files 5 and 6, and Supplemental Figures S2 and S3.

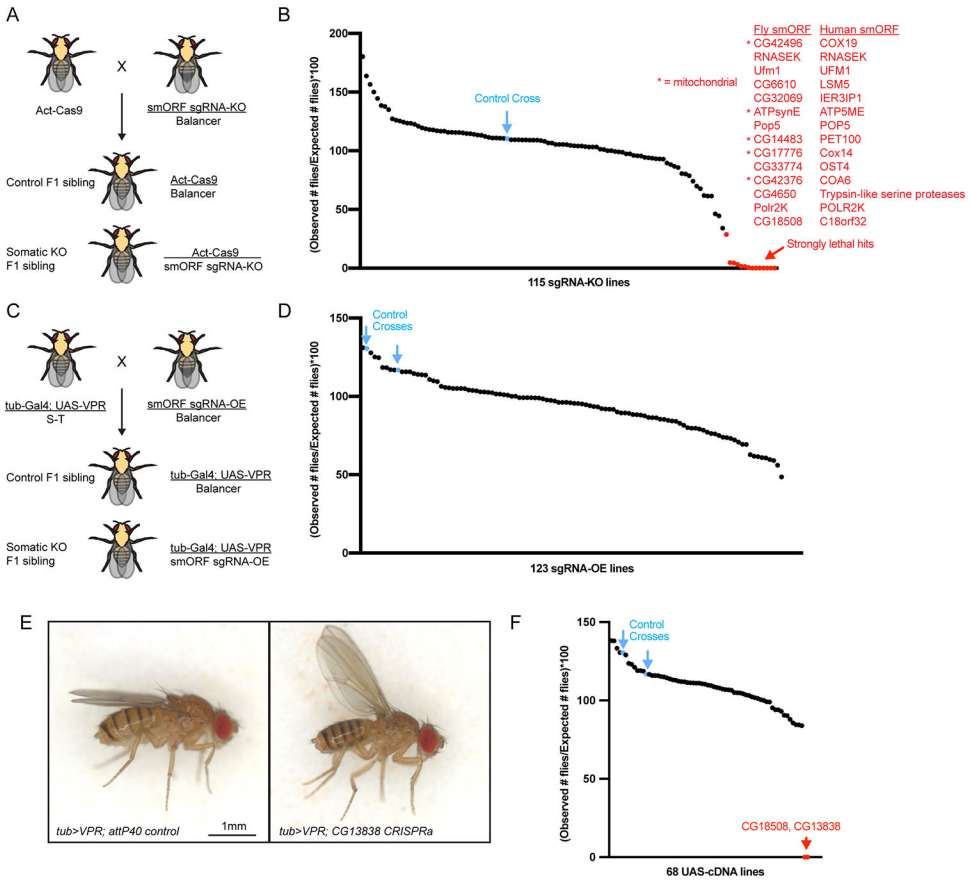


Figure 5. Functional characterization of conserved smORFs by F1 CRISPR in vivo screening.

(*A*) Genetic cross to perform CRISPR somatic knockout in F1 generation. **(*B*)** Quantification of viability of F1 flies from 115 sgRNA-KO crosses. Number of F1 progeny counted per cross was 918>n>33. **(*C*)** Genetic cross to perform CRISPR gene overexpression in F1 generation. ***(D)*** Quantification of viability of F1 flies from 123 sgRNA-OE crosses. Number of F1 progeny counted per cross was 220>n>56. **(E)** Images of adult female flies aged seven days after eclosion for two indicated genotypes. Scalebar = 1mm **(F)** Quantification of viability of F1 flies from 68 UAS-cDNA crosses. Number of F1 progeny countered per cross was 706>n>101. See also Supplemental File 7 and Supplemental Figures S4 and S5.

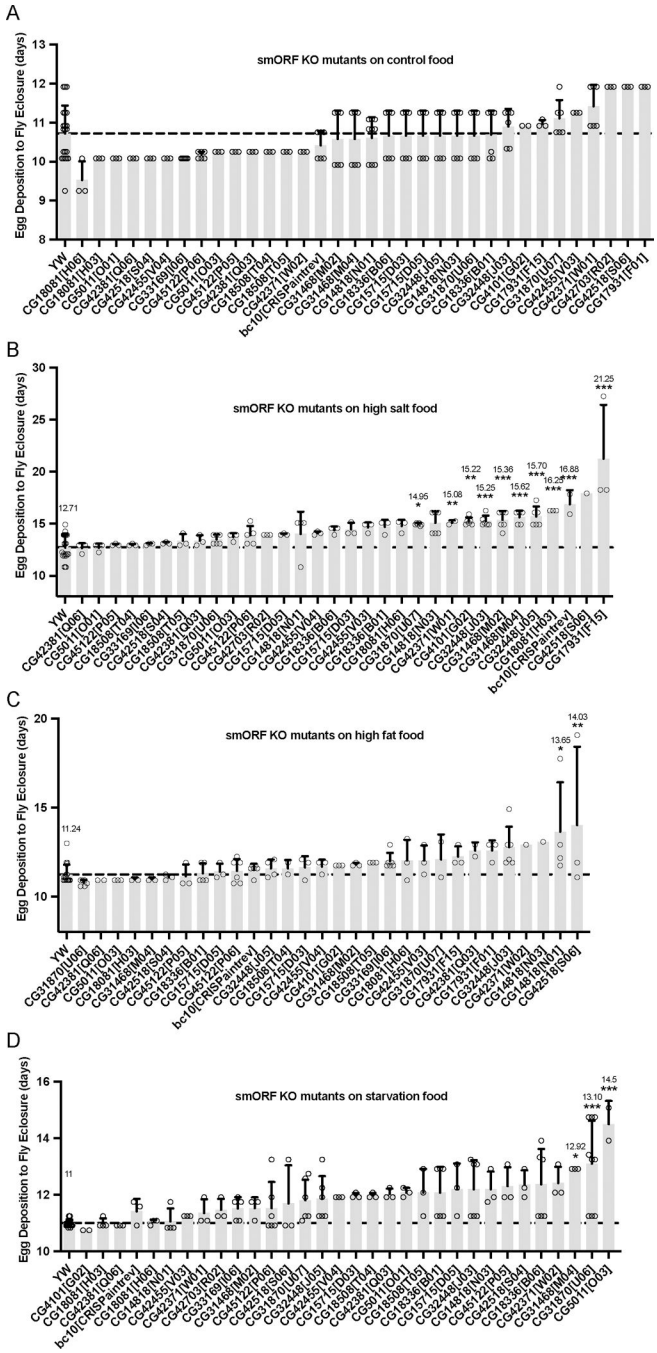


Figure 6. Full CRISPR knockout of 25 uncharacterized smORFs raised on stressful foods.

Developmental timing of egg deposition to adult eclosure in smORF KO mutants raised on ***(A)*** control food, ***(B)*** high salt food (30% NaCl), ***(C)*** high fat food (30% coconut oil), ***(D)*** starvation food (30% food, 70% PBS+1%agar). Significance was determined by One Way Anova test followed by a Dunnet Post-Hoc, P=0.05(*), P=0.01(**), P=0.001(***), error bars calculated with SD. See also Supplemental File 7 and Supplemental Figure S6. Each genotype-foodtype experiment was carried out with at least biological replicates.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
E.coli (TOP10)	Invitrogen	C404010
Chemicals, Peptides, and Recombinant Proteins		
Harringtonine	LKT Laboratories	H0169
Cycloheximide	Sigma Aldrich	C4859–1ML
BbsI	NEB	R3539
Critical Commercial Assays		
TRIzol Reagent	Thermo Fisher	15596026
RNeasy Mini Kit	QIAGEN	74106
NEBNext Ultra Directional RNA Library Prep Kit for Illumina	NEB	E7420
ZR small-RNA PAGE Recovery Kit	Zymo Research	R1070
NEBNext Small RNA Library Preparation Kit	NEB	E7330
Direct-zol RNA Miniprep kit	Zymo Research	R2050
iScript Reverse Transcription Supermix	BioRad	1708840
iQ SYBR Green Supermix	BioRad	170–8880
pENTR™/D-TOPO™ Cloning Kit	Invitrogen	K240020
LR Clonase II Enzyme mix	Invitrogen	11791–020
Gibson Assembly® Master Mix	NEB	E2611
Deposited Data		
RNA-seq data	NCBI Short Read Archive (SRA)	SRR18575339, SRR18575340, SRR18575342, SRR18575343, SRR18575345, SRR18575346
Experimental Models: Organisms/Strains		
Drosophila melanogaster	See Experimental Model and Subject Details and Supplemental File 7 for a list of fly strains	N/A
Oligonucleotides		
PCR primers	See Supplemental File 7	N/A
Recombinant DNA		
pCFD3	Port, F., Chen, H.M., Lee, T., and Bullock, S.L. (2014). Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proc Natl Acad Sci U S A *111*, E2967–2976. 10.1073/pnas.1405500111.	Addgene #49410
pCFD4	Port, F., Chen, H.M., Lee, T., and Bullock, S.L. (2014). Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proc Natl Acad Sci U S A *111*, E2967–2976. 10.1073/pnas.1405500111.	Addgene #49411

REAGENT or RESOURCE	SOURCE	IDENTIFIER
pCFD5	Port, F., and Bullock, S.L. (2016). Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs. Nat Methods *13*, 852–854. 10.1038/nmeth.3972.	Addgene #73914
pWalium10-roe	Perkins, L.A., Holderbaum, L., Tao, R., Hu, Y., Sopko, R., McCall, K., Yang-Zhou, D., Flockhart, I., Binari, R., Shim, H.S., et al. (2015). The Transgenic RNAi Project at Harvard Medical School: Resources and Validation. Genetics *201*, 843–852. 10.1534/genetics.115.180208.	DGRC: 1471
pCRISPaint-T2A-Gal4–3xP3-RFP	Bosch, J.A., Colbeth, R., Zirin, J., and Perrimon, N. (2020). Gene Knock-Ins in Drosophila Using Homology-Independent Insertion of Universal Donor Plasmids. Genetics *214*, 75–89. 10.1534/genetics.119.302819.	Addgene #127556
pGW-HA.attB	Bischof, J., Björklund, M., Furger, E., Schertel, C., Taipale, J., Basler, K. (2013). A versatile platform for creating a comprehensive UAS-ORFeome library in Drosophila. Development 140(11): 2434--2442. 10.1242/dev.088757	Bischof lab
pDONR223	Rual JF, Hirozane-Kishikawa T, Hao T, Bertin N, Li S, Dricot A, Li N, Rosenberg J, Lamesch P, Vidalain PO, Clingingsmith TR, Hartley JL, Esposito D, Cheo D, Moore T, Simmons B, Sequerra R, Bosak S, Doucette-Stamm L, Le Peuch C, Vandenhaute J, Cusick ME, Albala JS, Hill DE, Vidal M. Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res. 2004 Oct;14(10B):2128–35. doi: 10.1101/gr.2973604. PMID: 15489335	Vidal lab
Software and Algorithms		
STAR aligner v2.73a	http://code.google.com/p/rna-star/	RRID:SCR_004463
Picard v2.20.1	http://broadinstitute.github.io/picard/	RRID:SCR_006525
UCSC bedGraphToBigWig	https://genome.ucsc.edu/goldenPath/help/bigWig.html	N/A
Fastp v0.20.1	https://github.com/OpenGene/fastp	RRID:SCR_016962
FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	RRID:SCR_014583
Cutadapt	http://code.google.com/p/cutadapt/	RRID:SCR_011841
Bowtie2	https://github.com/BenLangmead/bowtie2	RRID:SCR_016368
ORFquant	https://github.com/lcalviell/ORFquant	N/A
Prism v9	http://www.graphpad.com/	RRID: SCR_002798
DELTA BLAST	Boratyn, G.M., Schäffer, A.A., Agarwala, R. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol Direct*7, 12 (2012). https://doi.org/10.1186/1745-6150-7-12	N/A
DIOPT	Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011 Aug 31;12:357. doi: 10.1186/1471-2105-12-357. PMID: 21880147; PMCID: PMC3179972	N/A