PLOS BIOLOGY

RESEARCH ARTICLE

# Functional unknomics: Systematic screening of conserved genes of unknown function

**João J. Rocha[1☯], Satish Arcot Jayaram[1☯], Tim J. Stevens[1☯], Nadine Muschalik[1☯], Rajen D. Shah[2], Sahar Emran[1], Cristina Robles[1], Matthew Freeman[1,3]*, Sean Munro[1]***

**1** MRC Laboratory of Molecular Biology, Cambridge, United Kingdom, **2** Centre for Mathematical Sciences, University of Cambridge, Cambridge, United Kingdom, **3** Sir William Dunn School of Pathology, University of Oxford, Oxford, United Kingdom

☯ These authors contributed equally to this work.
* matthew.freeman@path.ox.ac.uk (MF); sean@mrc-lmb.cam.ac.uk (SM)

Check for updates

OPEN ACCESS

**Citation:** Rocha JJ, Jayaram SA, Stevens TJ, Muschalik N, Shah RD, Emran S, et al. (2023) Functional unknomics: Systematic screening of conserved genes of unknown function. PLoS Biol 21(8): e3002222. https://doi.org/10.1371/journal.pbio.3002222

**Academic Editor:** Ian Dunham, European Bioinformatics Institute (EBI), UNITED KINGDOM

**Received:** January 12, 2023

**Accepted:** June 27, 2023

**Published:** August 8, 2023

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3002222

**Copyright:** © 2023 Rocha et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The Unknome can be viewed at http://unknome.org, with the entire database available to download as SQLite Version 3 files. Data from the functional screens are available

## Abstract

The human genome encodes approximately 20,000 proteins, many still uncharacterised. It has become clear that scientific research tends to focus on well-studied proteins, leading to a concern that poorly understood genes are unjustifiably neglected. To address this, we have developed a publicly available and customisable "Unknome database" that ranks proteins based on how little is known about them. We applied RNA interference (RNAi) in *Drosophila* to 260 unknown genes that are conserved between flies and humans. Knockdown of some genes resulted in loss of viability, and functional screening of the rest revealed hits for fertility, development, locomotion, protein quality control, and resilience to stress. CRISPR/Cas9 gene disruption validated a component of Notch signalling and 2 genes contributing to male fertility. Our work illustrates the importance of poorly understood genes, provides a resource to accelerate future research, and highlights a need to support database curation to ensure that misannotation does not erode our awareness of our own ignorance.

## Introduction

The advent of genome sequencing revealed in humans and other species thousands of open reading frames that encode proteins that had not been identified by earlier biochemical or genetic studies. Since the release of the first draft of the human genome sequence in 2000, the application of transcriptomics and proteomics has confirmed that most of these new proteins are expressed, and the function of many of them has been identified [1]. However, despite over 20 years of extensive effort, there are also many others that still have no known function [2,3]. The mystery and the potential biological significance of these unknown genes is enhanced by many of them being well conserved and often being unrelated to known proteins and thus lacking clues to their function. Analysis of publication trends has revealed that research efforts continue to focus on genes and proteins of known function, with similar trends seen in gene and protein annotation databases [2,4,5]. This is despite clear evidence from studies of gene expression and genetic variation that many of the poorly characterised proteins are linked to disease, including those that are eminently druggable [6,7]. Indeed, it has long been argued that ignorance can drive scientific advance [8].

in the main text or the supplementary data sets S2 and S3 Data. Code for the functional assays is available at https://github.com/tjs23/unknome.

**Funding:** This work was supported by the Medical Research Council, as part of United Kingdom Research and Innovation (MC_U105178783 to SM and MC_U105178780 to MF). Work in MF's lab was supported by Wellcome Investigator Awards 101035/Z/13/Z and 220887/Z/20/Z. RDS was funded by the Engineering and Physical Sciences Research Council (EP/R013381/1) and by the Alan Turing Institute through a Turing Fellowship (TU/B/00006). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** DUF, domain of unknown function; GO, Genome Ontology; GOA, Gene Ontology Annotation; MMAF, multiple morphological abnormalities of the sperm flagella; PCD, primary ciliary dyskinesia; RNAi, RNA interference; TRAP, translocon-associated protein; UPF, uncharacterised protein family.

This apparent bias in biological research toward the previously studied reflects several linked factors. Clearly, funding and peer-review systems are more likely to support research on proteins with prior evidence for functional or clinical importance, and individual perception of project risk seems likely to also contribute. In addition, scientific factors have been proposed, including a lack of specific reagents like antibodies or small molecule inhibitors, and a tendency to focus on proteins that are abundant and widely expressed and so likely to be present in cell lines and model organisms [4,7,9]. Finally, some genes may have roles that are not relevant to laboratory conditions [5].

Whatever the reasons, this inadvertent neglect of the unknown is clear and does not appear to be diminishing [9]. This has led to concern that important fundamental or clinical insight, as well as potential for therapeutic intervention, is being missed, and hence, the launch of several initiatives to address the problem. These include programmes to generate proteome-wide sets of reagents such as antibodies or mouse knock-out lines [10,11]. In addition, the NIH's Illuminating the Druggable Genome initiative supports work on understudied kinases, ion channels, and GPCRs [12]. There have been initiatives to develop new means to predict protein function or structure [13–17]. Finally, databases such as Pharos, Harmonizome, and neXtProt link human genes to expression and genetic association studies with the aim of highlighting understudied genes relevant to disease and drug discovery [18–20].

In this work, we have investigated directly the potential biological significance of conserved genes of unknown function by developing a systematic approach to their identification and characterisation. We have created an "Unknome database" that assigns to each protein from a particular organism a "knownness" score based on a user-controlled application of the widely-used Genome Ontology (GO) annotations [21,22]. The database allows selection of an "unknome" for humans, or a chosen model organism, that can be tuned to reflect the degree of conservation in other species, for example, allowing a focus on those proteins of unknown function that have orthologs in humans or are widely conserved in evolution. We use this database to evaluate the human unknome and find that it is shrinking only slowly. To assess the value of the unknome as a foundation for experimental work, we selected a set of 260 *Drosophila* proteins of unknown function that are conserved in humans and used RNA interference (RNAi) to test their contribution to a wide range of biological processes. This revealed proteins important for diverse biological roles, including cilia function and Notch pathway signalling. Overall, our approach demonstrates that significant and unexplored biology is encoded in the neglected parts of proteomes.

## Results

### Construction of an Unknome database

Much of the progress in understanding protein function has come from research in model organisms selected for their experimental tractability. Application of this research to the proteins of humans requires being able to identify the orthologs of these proteins in model organisms. Although it is not certain that orthologs in different species have precisely the same function, they generally have similar or related functions, implying that work from model organisms at the very least provides plausible hypotheses to test. Thus, our Unknome database was designed to link a particular protein with what is known about its orthologs in humans and popular model organisms.

A range of methods for identifying orthologs have been developed based on sequence conservation and although none are perfect, several achieve an accuracy in excess of 70%. We initially used the OrthoMCL database as it covered a wide range of organisms [23]. However, OrthoMCL was not being updated, and so the current Unknome database is based on the

PANTHER database (version 17.0) which covers over 143 organisms, is currently in continuous development, and has a good level of sensitivity and accuracy [24–26].

The heart of the Unknome database has been the development of an approach to assigning a "knownness" score to proteins. This is not trivial and is inevitably a somewhat subjective measure. Definitions of "known" range from a simple statement of activity to an understanding of mechanism at atomic resolution, and even well-characterised proteins can reveal unexpected extra roles. Thus, we designed the database so that the criteria for knownness can be user-defined, as well as having a default set of criteria. The GO Consortium provides annotations of protein function that are well suited to this application. Firstly, GO annotation is based on a controlled vocabulary and so is consistent between different species, and secondly, it is well structured thus allowing a user to apply their own definition of knownness.

The Unknome database combines PANTHER protein family groups (which we term "clusters") with the GO annotations for each member of the cluster. This includes annotations from humans and the 11 model organisms selected by the GO Consortium for their Reference Genome Annotation Project. The sequence-similar protein clusters (primary PANTHER families) not only contain orthologs, but also recent paralogs: duplications within individual species or lineages. The knownness score for each protein is calculated from the number of GO annotations it possesses.

It is important, however, to recognise that GO annotations do not all have equal evidential value, but they helpfully include an evidence code that indicates the type of source it is derived from. The Unknome database allows users to make use of this in generating a knownness score with an option to apply greater weight to annotations that are more likely to be reliable, such as those from a "Traceable Author Statement" rather than those "Inferred from Electronic Annotation" (Fig 1A and S1A Fig). In addition, weighting allows the selection of annotations most relevant to function. For instance, a protein's subcellular location is often included in its GO annotation, but this may not helpfully restrict the range of possible functions, so the database provides the option of excluding it when calculating a knownness value. The final knownness score of a cluster of proteins is set as the highest score of a protein in the cluster (Fig 1B).

The Unknome database is available as a website (http://unknome.org) that provides all protein clusters that contain at least 1 protein from humans or any of 11 model organisms (Fig 1C). The clusters can be ranked by knownness, and the user can modify this list so as to include only those proteins that are present in a particular combination of species, such as human plus a preferred model organism (Fig 1D). For each protein family, the interface shows the orthologs in its cluster and how the knownness of the cluster has changed over time (Fig 1C). These design principles maximise the versatility and power of the Unknome database as a tool for researchers from different biomedical fields.

## Validation of the Unknome database

To confirm that the Unknome database was accurately capturing current understanding of protein function, we ranked the 7,515 clusters of orthologs and paralogs that contain at least 1 human protein. Reassuringly, the top 10 scoring proteins have well-known roles in development and cell function (Fig 1E). In contrast, proteins containing one of the "Domains of Unknown Function" defined by the Pfam database were concentrated at the bottom of the range (S1B Fig). Clusters with a score of 1.0 or less correspond to 18.3% of all clusters but to 36% of the domains of unknown function (DUFs) and 59% of the related uncharacterised protein families (UPFs). The exceptions were typically multidomain proteins of known function that contain 1 domain whose role is unclear. Finally, the total number of PubMed citations for


**Fig 1. The Unknome database.** (A, B) Calculation of a knownness score for a cluster of orthologs based on the highest score in the cluster. Illustrated with a cluster corresponding to a subunit of a mitochondrial inner membrane translocase; (A) shows the GO annotations for mouse TIMM10, and derivation of a score based on the number of annotations weighted for their confidence, while (B) shows the scores for all the members of the cluster containing TIMM10 (UKP01389), with the highest score of a member being the knownness of the cluster. (C) The Unknome database contains information for each cluster showing its

distribution across species, links to information for the protein from each species, and the change in knownness over time—as illustrated for cluster UKP01389. (D) User interface to list clusters from a user-selected set of model organisms by the knownness of the cluster. The list indicates the best-known member of the cluster and the human member(s) of the cluster. (E) The 10 best known protein clusters, showing the best-known human gene in each. (F) Plot of the number of PubMed citations in the Uniprot comments section for human-gene containing clusters in the indicated range of knownness. The data underlying the plot can be found in S1 Data. GO, Genome Ontology.

https://doi.org/10.1371/journal.pbio.3002222.g001

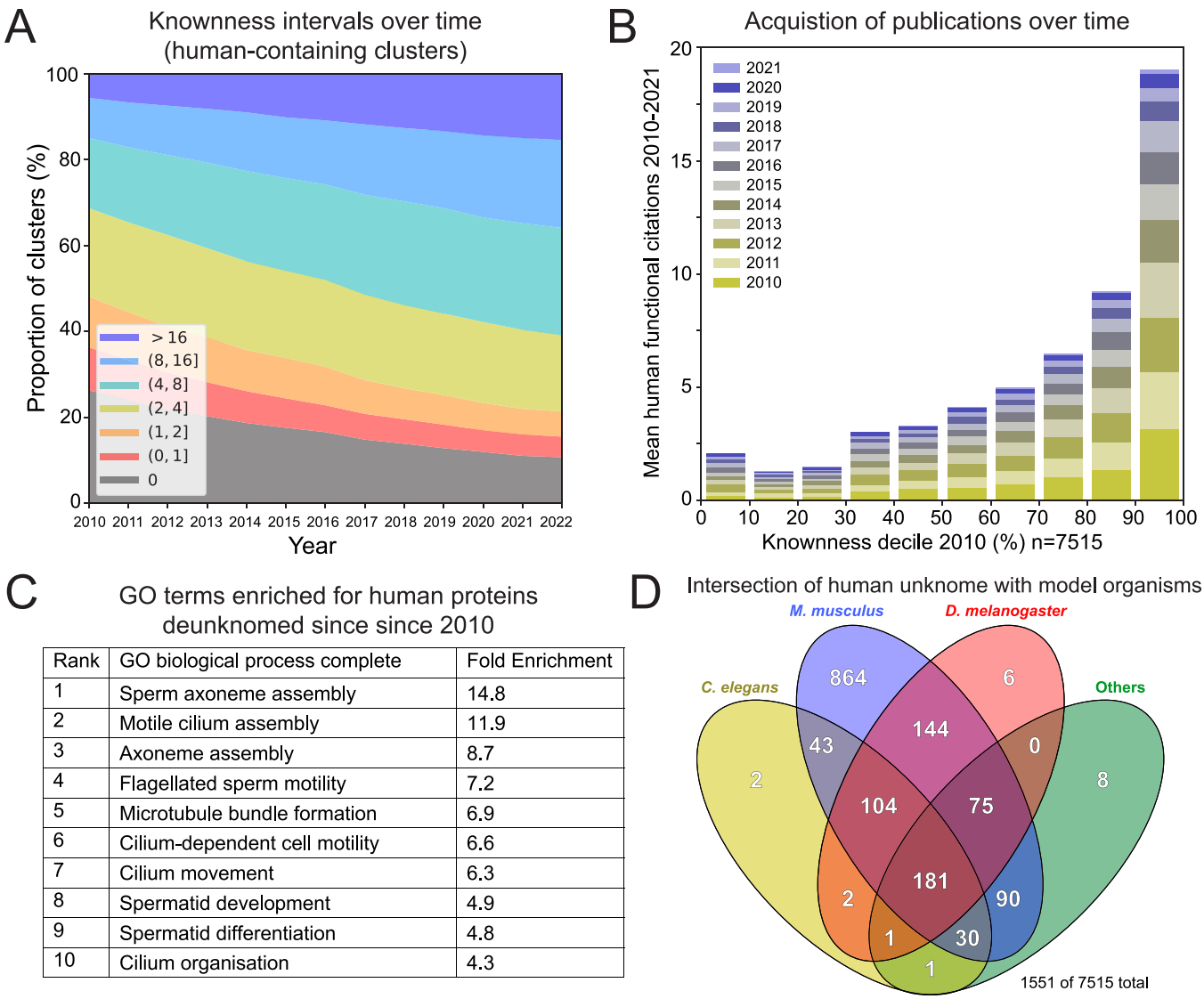
each protein shows a good correlation with the knownness scores from the database (Fig 1F). Overall, we conclude that the calculated knownness score provides a useful means to identify proteins of unknown function.

## The change of the Unknome over time

Unlike most databases, the Unknome will shrink over time. The knownness scores for clusters containing human proteins have increased across the whole range of proteins, but the proportion with a knownness score of 2 or less has declined from 43% to 23% over the last 10 years, with the decline being less in nonhuman model organisms (Fig 2A and S2A Fig). This slow progress is unlikely to represent a deficit in GO annotation which is kept up to date, but rather that human genes and proteins are much more likely to have been published on in the last 12 years if they are in clusters that were already well known at the start of this period (Fig 2B and S2B Fig). Consistent with this, knownness increases more rapidly over time for genes that were already well annotated (S2C Fig). These observations provide further support to the notion that research activity tends to focus on what has already been studied in depth [2,4,27]. There are 750 human clusters whose knownness was zero 12 years ago but has since increased to above 2. The GO terms most enriched in this set are mostly associated with cilia, reflecting recent acceleration of progress in studying this large and complex structure that is absent from some model organisms such as yeast (Fig 2C). Consistent with this, the less known human genes tend to be less likely to be conserved outside of vertebrates, and generally have fewer orthologs, suggesting that progress has been hampered by there being fewer orthologs that could be found by genetic screens in non-vertebrates (S2D and S2E Fig). Interestingly, the most highly known proteins are also less likely to be conserved outside of metazoans, reflecting the fact that many are involved in important developmental pathways or signalling events relevant to multicellularity (S2D Fig). However, of the 1,606 human-containing clusters with a current knownness score of less than 2.0, 68% are detectably conserved outside of vertebrates and 45% are conserved outside of metazoans (Fig 2D). Interestingly, no one model organism contains all of these, indicating that each has a role to play in illuminating the human unknome.

## Functional unknomics in *Drosophila*

To test the value of the Unknome database, and to pilot experimental approaches to studying neglected but well-conserved proteins, we selected a set of unknown human proteins that are conserved in *Drosophila* and hence amenable to genetic analysis. *Drosophila* also tends to lack partial redundancy between closely related paralogs, as in humans this arose in many gene families from the 2 whole-genome duplications that occurred early in vertebrate evolution [28]. A powerful approach to investigating gene function in *Drosophila* is to knockdown its expression with RNAi and assess the biological consequences [29,30]. We thus determined the effect of expressing hairpin RNAs to direct RNAi against a panel of genes of unknown function.




| Rank | GO biological process complete | Fold Enrichment |
|---|---|---|
| 1 | Sperm axoneme assembly | 14.8 |
| 2 | Motile cilium assembly | 11.9 |
| 3 | Axoneme assembly | 8.7 |
| 4 | Flagellated sperm motility | 7.2 |
| 5 | Microtubule bundle formation | 6.9 |
| 6 | Cilium-dependent cell motility | 6.6 |
| 7 | Cilium movement | 6.3 |
| 8 | Spermatid development | 4.9 |
| 9 | Spermatid differentiation | 4.8 |
| 10 | Cilium organisation | 4.3 |

**Fig 2. Analysis of trends in knownness.** (A) Change in the distribution of knownness of the 7,515 clusters that contain at least 1 protein from humans. (B) Mean number of publications added each year since 2010 to the UniProt entry for the human protein in each of the 7,515 clusters that contain at least 1 human protein, ranked into deciles based on knownness at 2010. Where there was more than 1 human protein in the cluster, their publications were summed. The best-known clusters in 2010 received the most publications in subsequent years. (C) The 10 largest GO term enrichments for the 753 human proteins from clusters whose knownness has increased from 0 in 2010 to 2.0 or above by 2022. When there was more than 1 human protein in the cluster, a single one was used chosen by alphabetical order to avoid bias. GO enrichment analysis used ShinyGO [112]. (D) Venn diagram showing the distribution of genes from the indicated species in the 1,551 clusters of knownness <2.0 and which contain at least 1 human protein. Not shown are the 55 clusters that appear only in humans. The data underlying the graphs shown in the figure can be found in S1 Data. GO, Genome Ontology.

https://doi.org/10.1371/journal.pbio.3002222.g002

We initially selected all genes that had a knownness score of ≤1.0 and are conserved in both humans and flies, as well as being present in at least 80% of available metazoan genome sequences. Of the 629 corresponding *Drosophila* genes, 358 were available in the KK library that was the best available genome-wide RNAi library at the time (S1 Table) [31]. This, and other RNAi libraries, have been used for several genome-wide screens for phenotypes readily analysed at large scale, but had not been used for the screens that we applied [31]. These KK library stocks were crossed to lines containing Gal4 drivers to express the hairpin RNAs in

either the whole fly or in specific tissues. After testing for viability, the nonessential genes were then screened with a panel of quantitative assays designed to reveal potential roles in a wide range of biological functions. These include male and female fertility, tissue growth (in the wing), response to the stresses of starvation or reactive oxygen species, proteostasis, and locomotion. The results of these screens are discussed below.

## Unknown genes have essential functions

To determine if the genes were required for viability, a ubiquitous GAL4 driver was used to direct RNAi throughout development (daughterless-Gal4). For 162 of the 358 genes, the resulting progeny showed compromised viability with either all (lethal) or almost all (semi-lethal) failing to develop beyond pupal eclosion, suggesting that these genes are essential for development or cell function (S1 Table). However, it was subsequently reported that in a subset of the lines in the KK RNAi library, the transgene is integrated in a locus (40D) that itself results in serious developmental defects when the transgene is expressed with a GAL4 driver [32,33]. Following PCR screening, we removed all of the stocks that had this integration site, all but one of them having been lethal in the initial screen. For the remaining 260 genes, the stocks used the alternative integration site which is not problematic, with KK stocks having been used successfully in a range of different screens [29,34]. For these, the RNAi compromised viability in 62 cases (24%). In considering the results from RNAi screens, one must always be mindful of off-target effects, and in *Drosophila*, the possible effects of variability in genetic background and conditions of rearing and maintenance. Nonetheless, of these 62 genes, 12% were also identified in a recent genome-wide screen of genes required for viability of S2 cells; in contrast, only 4% of the 198 nonessential genes were hits in the S2 cell screen [35]. The S2 study estimated that 17% of genes known to be essential in flies are also essential in S2 cells, and it is likely that using RNAi to knockdown gene function underestimates lethality. Our screen in whole organisms reveals that, despite several decades of extensive genetic screens in *Drosophila*, there are many genes with essential roles that have eluded characterisation.

Of course, there is more to life than being alive. We therefore subjected the 198 apparently nonessential genes to a range of phenotypic tests to determine if they had detectable roles in a wide range of organismal functions. On the grounds that the long history of *Drosophila* genetic screens may have saturated the discovery of mutants with easily detectable phenotypes (mostly developmental defects), we targeted our search to nonstandard and quantitative phenotypes that are harder to assess. In practice, this meant designing phenotypic screens that were more complex than normal. Our hope was that this would identify a larger proportion of genes that had not been hit in more standard *Drosophila* screens. The results of these function screens are described below, followed by a validation of selected hits, with the screening data provided in S2 and S3 Data and the results summarised in S2 Table.

## Contribution of unknome genes to fertility

To test fertility, specific GAL4 drivers were used to knockdown the set of 198 unknown genes in either the male or female germline. Even with collecting data for multiple flies per gene, the resulting brood sizes showed some variability, as expected for a quantitative measure of a biological process. Thus, for all our assays, we needed to determine if outliers had a phenotype that exceeded to a statistically significant degree the variation intrinsic in the population. To do this, we used statistical tests based on 3 steps. First, we performed a regression on the replicate data for each gene to estimate its parameters and standard errors within the assay. Next, an outlier region was determined by fitting the parameter estimates for all analysed genes to a normal distribution, which was then used to define a boundary for outliers. Finally, for each

gene, we tested the hypothesis that it falls within the outlier boundary. This approach is summarised in the Methods and described in detail in the Supporting information (S1 Text). To display the data from the fertility tests, mean brood sizes obtained from RNAi-treated males was plotted against those obtained from RNAi-treated females for each gene (Fig 3A). Several of the RNAi lines gave a substantial reduction in brood size that was sex specific and highly statistically significant.

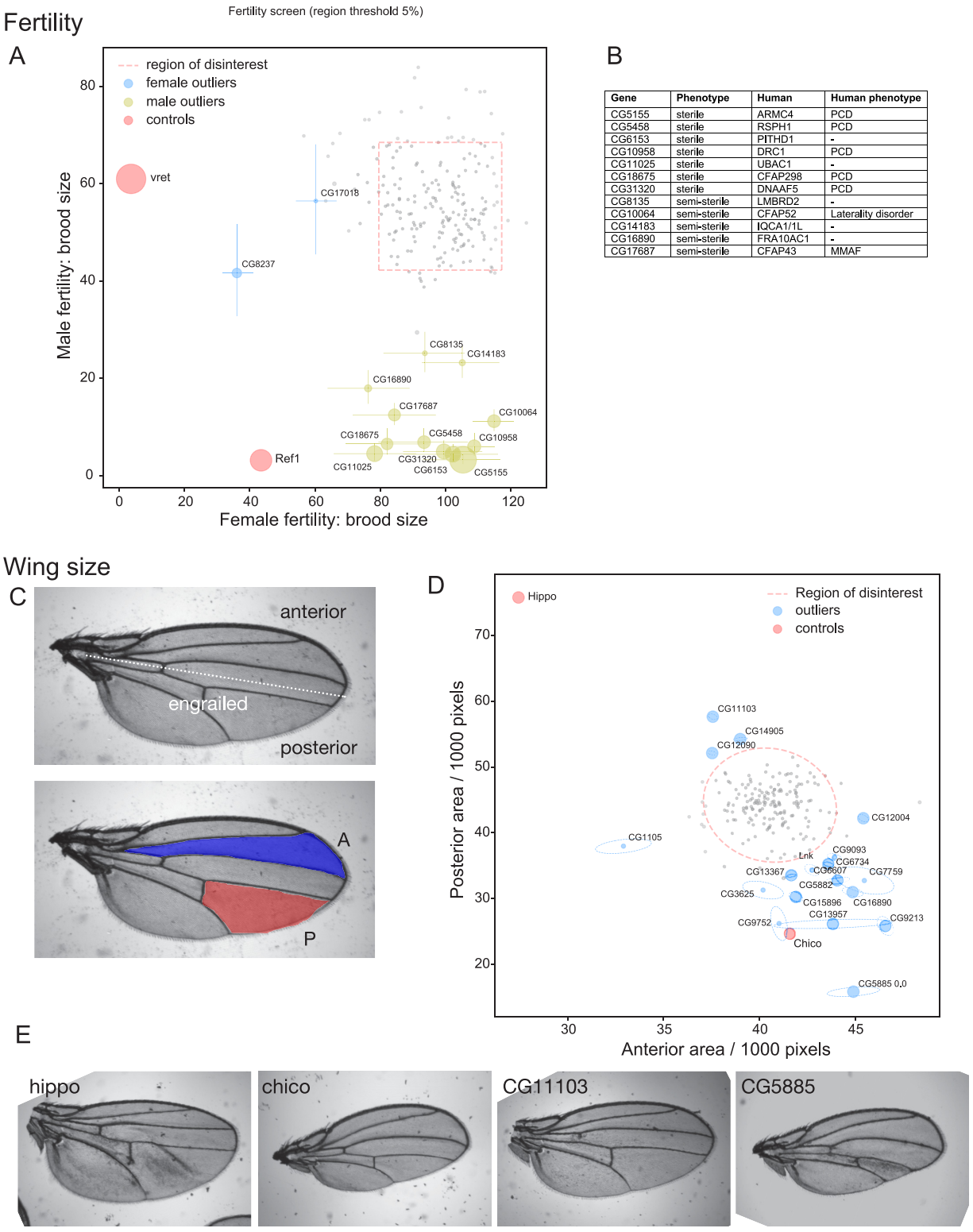
**Female fertility.** Two genes gave a partial, but significant, reduction in female brood size. During the course of our work, a mouse ortholog, MARF1, of one of these hits, CG17018, was identified in a genetic screen as being required for maintaining female fertility, apparently by controlling mRNA homeostasis in oocytes [36,37]. A recent study of CG17018 has confirmed that it is indeed required for female fertility in *Drosophila*, despite lacking some domains present in MARF1. Its appearance as a hit in our screen is therefore an encouraging validation of the approach [38]. The other gene, CG8237, has not previously been linked to fertility, but has a mammalian ortholog (FAM8A1) that has been recently proposed to help assemble the machinery for ER-associated degradation (ERAD) and so may have an indirect effect on oogenesis [39,40]. We selected CG8237 for validation by CRISPR/Cas9 gene disruption as described below.

**Male fertility.** Seven genes showed near complete male sterility, with 5 further genes giving a statistically significant reduction in brood size. In humans, male sterility is one of the symptoms associated with primary ciliary dyskinesia (PCD), a disorder affecting motile cilia and flagella. While our analysis was in progress, exome-sequencing allowed the identification of many new PCD genes [41,42]. Interestingly, 5 of the genes identified in our assay are homologs of human PCD genes (Fig 3B), of which CG5155 (ARMC4) and CG31320 (DNAAF5) have since been shown to be required in *Drosophila* for male fertility [43,44]. All of these genes comprise, or help assemble, the dynein-based system that drives the beating of cilia and flagella. In addition, human orthologs of 2 of the semi-sterile hits in the Unknome screen have been found to be mutated in related familial conditions. CFAP43 (orthologous to CG17687) is mutated in patients with multiple morphological abnormalities of the sperm flagella (MMAF), and CFAP52 (orthologous to CG10064) is mutated in laterality disorder, a condition caused by defects in ciliary beating during development [45,46]. A further semi-sterile hit, CG14183, is an ortholog of DRC11, a subunit of the nexin-dynein regulatory complex that regulates flagellar beating in *Chlamydomonas* [47]. These findings prove the value of the Unknome database approach to identifying new genes of biological significance and validate the RNAi-based screening approach.

Of the 4 remaining genes that showed male fertility defects, CG11025 is now only partially unknown as its human ortholog (UBAC1) is a non-catalytic subunit of the Kip1 ubiquitination-promoting complex, an E3 ubiquitin ligase [48]. CG11025 was recently identified in a genetic screen for defects in ciliary traffic and found to be required for fertility [49]. However, the other 3 genes, CG8135, CG6153, and CG16890 (orthologous to LMBRD2, PITHD1, and FRA10AC1), remain poorly understood in any species. They are less likely to be flagellar components as they are not predominantly expressed in testes and, as described below, 2 were selected for validation by CRISPR/Cas9 gene disruption, along with CG10064 whose ortholog CFAP52 is mutated in laterality disorder.

## Contribution of unknome genes to tissue growth

To test the unknome set of genes for roles in tissue formation and growth, we examined the effect of knocking them down in the posterior compartment of the wing imaginal disc and comparing the area of the posterior compartment of the adult wing to that of the control




| Gene | Phenotype | Human | Human phenotype |
|---|---|---|---|
| CG5155 | sterile | ARMC4 | PCD |
| CG5458 | sterile | RSPH1 | PCD |
| CG6153 | sterile | PITHD1 | - |
| CG10958 | sterile | DRC1 | PCD |
| CG11025 | sterile | UBAC1 | - |
| CG18675 | sterile | CFAP298 | PCD |
| CG31320 | sterile | DNAAF5 | PCD |
| CG8135 | semi-sterile | LMBRD2 | - |
| CG10064 | semi-sterile | CFAP52 | Laterality disorder |
| CG14183 | semi-sterile | IQCA1/1L | - |
| CG16890 | semi-sterile | FRA10AC1 | - |
| CG17687 | semi-sterile | CFAP43 | MMAF |

**Fig 3. Testing of the unknome set of genes for roles in fertility and wing growth.** (A) Plot of brood sizes obtained from matings in which each gene was knocked down in either the male or female germline. Dotted lines indicate outlier boundaries, with the genes named being those whose position outside of the boundary is statistically significant, error bars show standard deviation, and the size of the circles is inversely proportional to the *p*-value. Controls: Vret is involved in piRNA biogenesis and affects female fertility [113], and Ref1 is an essential protein predicted to be involved in RNA export [114], and affects both males and females. (B) Summary of the significant hits from the test of male

fertility, showing the human ortholog and the phenotype reported for patients with loss of function mutations (PCD, MMAF). (C) Adult wing illustrating the posterior domain that expresses engrailed during development and hence the engrailed-Gal4 driver used to express the hairpin RNAs. Also shown are the intervein areas measured to assess tissue growth in the anterior and posterior halves of the wing. (D) Plot of the mean area of the anterior and posterior intervein areas as in (C) for flies in which each gene was knocked down by RNAi in the posterior domain (pixel dimensions 2.5 μm × 2.5 μm). Errors are shown as tilted ellipses with the major/minor axes being the square roots of the eigenvectors of the covariance matrix. Dotted lines indicate the outlier boundary, with the genes named being those whose position outside of the boundary is statistically significant, with the size of the circles being inversely proportional to the *p*-value. The genes Hippo (growth repressor) and Chico (growth stimulator) were included as controls. (E) Representative wings from flies expressing hairpin RNA for the indicated genes in the posterior domain. Hippo and Chico are controls as in (D), with CG11103 and CG5885 showing an increase or decrease in the posterior domain, respectively. The means and variances used for the graphs shown in the figure can be found in S2 Data with the data points in S3 Data. MMAF, multiple morphological abnormalities of the sperm flagella; PCD, primary ciliary dyskinesia; RNAi, RNA interference.

https://doi.org/10.1371/journal.pbio.3002222.g003

anterior compartment (Fig 3C), a method previously used to detect effects of a range of different genes [50,51]. As controls, we used Hippo, a negative regulator of tissue size, and Chico, a component of the PI 3-kinase pathway that stimulates organ growth [52,53]. Knockdown of 3 of the unknome genes in the posterior compartment caused a statistically significant increase in its area (Fig 3D and 3E). These include CG12090, the *Drosophila* ortholog of mammalian DEPDC5, which was found to be part of the GATOR1 complex that inhibits the Tor pathway during the protracted course of our studies. Mutants in GATOR1 subunits promote cell growth by increasing Tor activity [54,55]. The other 2 are CG14905 and CG11103. CG14905 is a paralog of a testes-specific gene CG17083, and both are orthologs of mammalian CCDC63/CCDC114 that have a role in attaching dynein to motile cilia, although CG14905 seems likely to have additional roles as it is ubiquitously expressed [56]. CG11103 (TM2D2) encodes a small membrane protein that shares a TM2 domain with Almondex, a protein with an uncharacterised role in Notch signalling [57]. We therefore selected CG11103 for further validation by CRISPR/Cas9 as described below.

A larger number of genes caused a reduced compartment size when knocked down (Fig 3D). However, this could arise from a wide range of causes and so this is broad ranging assay for protein importance, and indeed mammalian orthologs of several of the stronger hits have been subsequently found to act in known cellular processes such membrane traffic (CG13957, the ortholog of human WASHC4), lipid degradation (CG3625/AIG1), or tRNA production (CG15896/PRORP). The strongest effect was seen with CG5885, an ortholog of a subunit of the translocon-associated protein (TRAP) complex that is associated with the Sec61 ER translocon [58]. TRAP's role is enigmatic and so it was also selected for CRISPR/Cas9 validation.

## Contribution of unknome genes to protein quality control

The removal of aberrant proteins is a fundamental aspect of cellular metabolism, and thereby organismal health, but it is a function that does not necessarily contribute substantially to well-screened developmental phenotypes. It also exemplifies our suspicion that a disproportionately high number of the unknome set of genes may be involved in quality control and stress response functions, which are likely to have been missed by many traditional experimental approaches. We therefore tested the unknome gene set for protein quality control phenotypes, using an assay based on aggregation of GFP-tagged polyglutamine, a structure found in mutants of huntingtin that cause Huntington's disease [59]. When this Httex1-Q46-eGFP reporter is expressed in the eye, the aggregates can be detected by fluorescence imaging (Fig 4A). The RNAi guides were co-expressed in the eye to knockdown unknome genes, and the number of polyQ aggregates quantified for 2 different size ranges. Although there was considerable variation in aggregate number, statistical analysis allowed the identification of clear outliers among the unknome RNAi set (Fig 4B). Most of the genes showing the largest increase

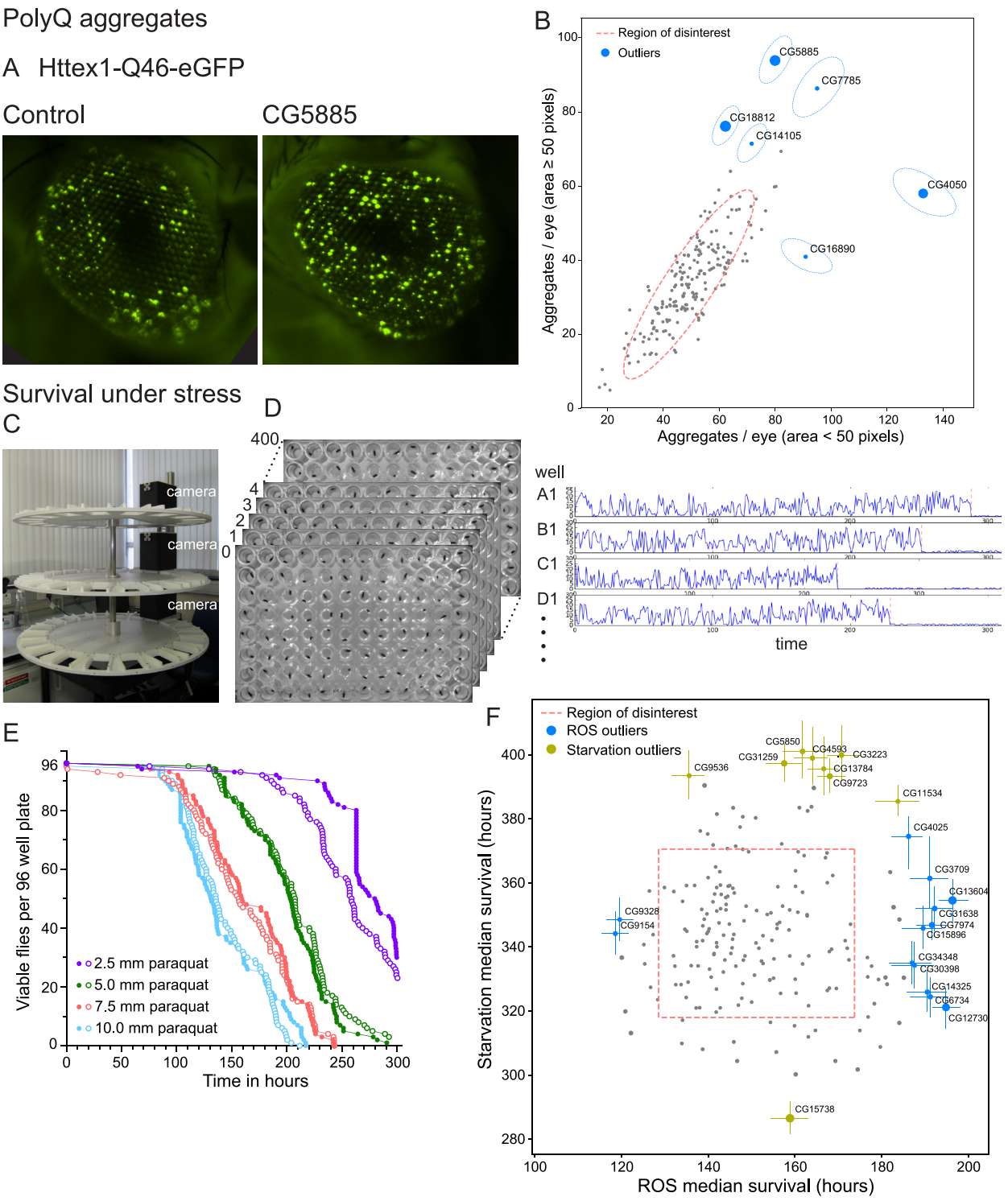



**Fig 4. Testing of the unknome set of genes for roles in quality control and responses to stress.** (A) Fluorescence micrographs of eyes from stocks expressing Httex1-Q46-eGFP along with either no RNAi, or one to the screen hit CG5885, both under the control of the GMR-GAL4 driver. The GFP fusion protein forms aggregates whose number and size increase over time. (B) Plot of the mean number of large (≥50 pixels) or small (<50 pixels) aggregates of Httex1-Q46-eGFP formed after 18 days in flies in which the unknome set of genes has been knocked-down by RNAi (pixel dimensions 0.5 μm × 0.5 μm). Errors are shown as tilted ellipses with the major/minor axes being the square roots of the eigenvectors of the covariance matrix. Dotted lines indicate an outlier boundary set at 90% of the variation in the dataset, with the genes named being those whose position outside of the boundary is statistically significant with a *p*-value <0.05, with the size of the circles being inversely proportional to the *p*-

value. (C) Flywheel apparatus for time-lapse imaging of 96-well plates containing 1 fly per well. Each of 3 wheels holds 20 plates that rotate under a camera to be imaged once per hour. (D) Use of time-lapse imaging to assay viability: 96-well plates were imaged very hour and the movement between frames quantified for the fly in each well. Plots of movement size over time allow the time point for cessation of movement and hence loss of viability to be determined automatically. (E) Survival plots obtained from the flywheel for flies in 96-well plates with food containing the indicated concentration of oxidative stressor paraquat. Increased levels of the paraquat shorten survival times. Two independent 96-well plates are shown for each condition to illustrate the reproducibility of the assay. (F) Plot of the median survival time of fly lines in which the unknome set of genes has been knocked-down by RNAi and which were then exposed to paraquat to induce oxidative stress or were starved for amino acids. Dotted lines indicate an outlier boundary set at 80% of the variation in the dataset, with the genes named being those whose position outside of the boundary is statistically significant ($p$-value $<0.05$), with error bars showing standard deviation and the size of the circles inversely proportional to the $p$-value. The means and variances used for the graphs shown in (B) and (F) can be found in S2 Data with the individual data points in S3 Data. The data underlying the graph in (E) can be found in S1 Data. RNAi, RNA interference.

https://doi.org/10.1371/journal.pbio.3002222.g004

in aggregates remain of unknown function (CG7785 (SPRYD7 in humans), CG16890 (FRA10AC1), CG14105 (TTC36), and CG18812 (GDAP2)), although mutation of GDAP2 in humans causes neurodegeneration, consistent with a role in quality control [60]. More is now known about 2 of the hits. CG4050 is a mammalian ortholog of TMTC3, one of a family of ER proteins recently shown to be O-mannosyltransferases; deletion of TMTC3 causes neurological defects [61,62]. CG5885 is the ortholog of the SSR3 subunit of the TRAP complex that also showed reduced wing size; in mammalian cells, the TRAP complex is up-regulated by ER stress [58]. These hits are consistent with reports that ER stress can increase cytosolic protein aggregation [63].

## Contribution of unknome genes to resilience to stress

Genomes have evolved to deal with many environmental stresses, and again, these are processes poorly investigated by traditional genetic approaches. We therefore tested resilience to stress, following knockdown of the unknome set. To quantify the viability of large numbers of flies, individual flies were arrayed in 96-well plates, and the plates maintained on a "flywheel" that rotated them under a camera every hour (Fig 4C and S1 Video). Viability was indicated by movement between images, allowing time of death to be determined with an accuracy of +/− 1 h (Fig 4D and 4E). We applied this method with 2 challenges likely to be associated with different cellular resilience mechanisms: amino acid starvation and oxidative stress.

**Resilience under starvation.** Under conditions of amino acid deprivation, knockdown of 8 of the unknome test set significantly prolonged survival (Fig 4F). Seven of these genes remain of unknown function, but interestingly, 5 have orthologs in other species whose localisation or interactions suggest that they have roles in the endosomal system. Thus DEF8, the mammalian ortholog of CG11534, has been reported to interact with Rab7 [64,65], and TMEM184A (CG5850) has been reported to act in the endocytosis of heparin [66]. In addition, the mammalian orthologs of CG4593 and CG9536 (CCDC25 and TMEM115) are Golgi-localised proteins of unknown function, and the yeast ortholog of CG13784 (ANY1) has been found to suppress loss of lipid flippases that act in endosome-to-Golgi recycling [67,68]. Our identification of this cluster of genes with related functions suggests that defects in endocytic recycling can prolong survival in starvation, possibly by altering autophagy or by reducing signalling from receptors that promote anabolism. The other 2 genes that improved starvation resilience when knocked down have no known function in any species, with loss of CG31259 (TMEM135) causing mitochondrial defects, and nothing reported for CG3223 (UBL7) [69,70]. One gene, CG15738 (NDUFAF6), caused an increased susceptibility to starvation, and it has been found to be an assembly factor for mitochondrial complex I, whose loss compromises viability [71].

**Resilience under oxidative stress.** Resistance to oxidative stress was tested with paraquat, an insecticide widely used to elevate superoxide levels in *Drosophila* [72,73]. There was considerable variability in the survival times, but 11 genes gave a statistically meaningful increase in resistance (Fig 4F). Most of these genes remain unknown, but 3 have since been reported to have functions related to oxidative stress signalling. The mammalian ortholog of CG4025 (DRAM1/2) is induced by p53 in response to DNA damage and promotes apoptosis and autophagy [74]. The mammalian orthologs of CG13604 (UBASH3A/B) are tyrosine phosphatases that repress SYK kinase, an enzyme reported to help protect cells against ROS, with superoxide activation of *Drosophila* Syk kinase signalling tissue injury [75–77]. Finally, the ortholog of CG3709 in archaea has tRNA pseudouridine synthase activity, but the human ortholog PUS10 has been reported to be cleaved during apoptosis and promote caspase-3 activity, thus its loss may slow apopotic cell death [78]. Of the other 8 hits, 5 remain poorly characterised, 1 is involved in mitochondrial function and so may reduce ROS production, and 2 are involved microtubule function with no clear link to superoxide responses. Although further validation will be required, these 5 genes seem good candidates to have a role in mitochondria or ROS-response pathways.

## Contribution of unknome genes to locomotion

Metazoans benefit from having a musculature under neuronal control. We therefore addressed the possibility of neuromuscular functions by testing the role of the unknome set of genes in locomotion, using the iFly tracking system in which the climbing trajectories of adult flies are quantified by imaging and automated analysis (Fig 5A) [79,80]. Climbing speed declines with age, so the assay was performed at both 8 days and 22 days post eclosion. Climbing speeds are inevitably somewhat variable, even in wild-type flies, but nonetheless 6 genes were statistically significant outliers when assayed after 8 days (Fig 5B). Two of these genes remain poorly understood, and for 3 of the others recent work indicates a role in muscle or neuronal function. These include CG9951, whose human homolog CDCC22 has been recently found to be a subunit of the retriever complex that acts in endosomal transport. Missense mutations in CDCC22 causing intellectual disability [81,82]. The human ortholog of CG13920 (TMEM35A) is required for assembly of acetylcholine receptors [83]. Finally, CG3479 is the gene mutated in the *Drosophila outspread (osp)* wing morphology allele, and is expressed in muscle, with one of its 2 mammalian orthologs (MPRIP) being been found to regulate actinomyosin filaments [84,85].

## Validation of fertility screen hits by gene disruption

Analysis of gene function by RNAi can be confounded by off-target effects. We therefore used CRISPR/Cas9 gene disruption to validate selected hits from 2 of the phenotypic screens. From the fertility screens, 3 male steriles and 1 female sterile were selected for genetic disruption. Of the male hits, CG10064 and CG6153 were both confirmed as being required for male fertility (Fig 6A to 6D). CG10064 is a WD40 repeat protein, and mutation of its human ortholog, CFAP52, results in abnormal left-right asymmetry patterning, a process known to depend on motile cilia [46,86]. CG6153 comprises a PITH domain that is also found in TXNL1, a thioredoxin-like protein that associates with the 19S regulatory domain of the proteasome through its PITH domain [87,88]. Males lacking CG6153 made morphologically normal sperm, but they did not accumulate in the seminal vesicle, the organ in which nascent sperm are stored prior to deployment, suggesting that they have limited viability (Fig 6E to 6J). Neither CG6153 nor its human ortholog PITHD1 are testis specific, and, indeed, orthologs are also present in non-ciliated plants and yeasts, suggesting that the protein has a role in an aspect of proteasome

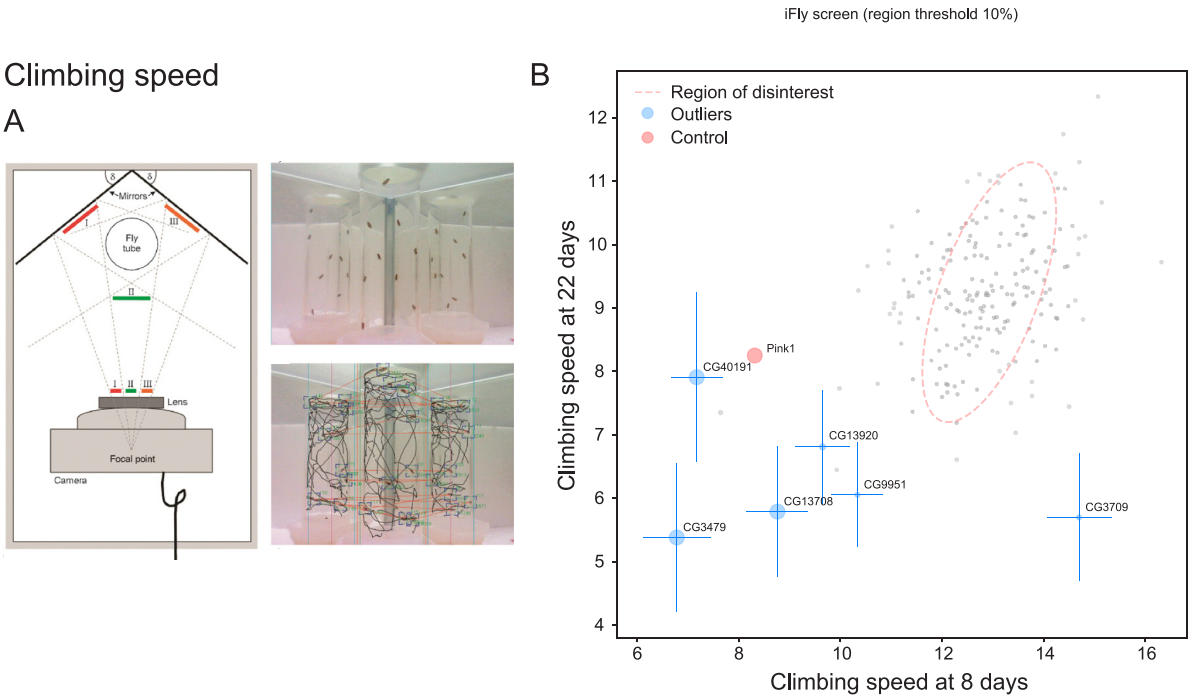



**Fig 5. Testing the unknome set of genes for roles in locomotion.** (A) iFly tracking system for automatic quantitation of *Drosophila* locomotion (reproduced from Kohlhoff and colleagues [80]). *Drosophila* are knocked to the bottom of a glass vial and placed in an imaging chamber that allows viewing from 3 angles and their climbing tracked automatically. (B) Plot of the mean climbing speeds of fly lines in which the unknome set of genes has been knocked down by RNAi, and the speeds for each line were determined after 8 days or 22 days post eclosion. Loss of the Parkinson's gene Pink1 affects climbing speed and it was included as a control [115]. Dotted lines indicate an outlier boundary set at 90% of the variation in the dataset, with the genes named being those whose position outside of the boundary is statistically significant with a $p$-value $<0.1$, with error bars showing standard deviation and the size of the circles inversely proportional to the $p$-value. The means and variances used for the plot shown in the figure can be found in S2 Data with the data points in S3 Data. RNAi, RNA interference.

https://doi.org/10.1371/journal.pbio.3002222.g005

biology that is of particular importance for maturing viable sperm. Recent work on mouse PITHD1 indicates it has a role in both olfaction and fertility [89,90]. The other male sterile hit, CG16890 (FRA10AC1), and the female sterile hit, CG8237 (FAM8A1), did not show reduced fertility when disrupted and presumably represent off-target RNAi effects (S3 Fig).

## Wing size hit CG11103 is a regulator of Notch signalling

Knockdown by RNAi of gene CG11103 (TM2D2 in humans) caused alterations in the growth of the wing (Fig 3D and 3E). When CG11103 was removed with CRISPR/Cas9, mutant females and males were viable without any obvious phenotypes, but females were completely sterile (Fig 7A and 7B). Eggs laid by mutant females were fertilised but failed to develop, and cuticle preparations and antibody labelling of the pan-neuronal marker Elav showed a hyperplasia of nervous system at the expense of the epidermis (Fig 7C–7G). This phenotype is characteristic of defects in the highly conserved Notch signalling pathway that is required in the *Drosophila* embryo to determine/specify the neuroblasts that give rise to the CNS in a process called lateral inhibition. CG11103 contains a TM2 domain that comprises 2 putative transmembrane domains connected by a short linker [91]. The function of this domain is unknown, but it occurs in 2 related proteins in *Drosophila*, and all 3 of the fly proteins have human orthologs (Fig 7B). Interestingly, one of these, *almondex*/*CG12127*, was identified as a gene required for

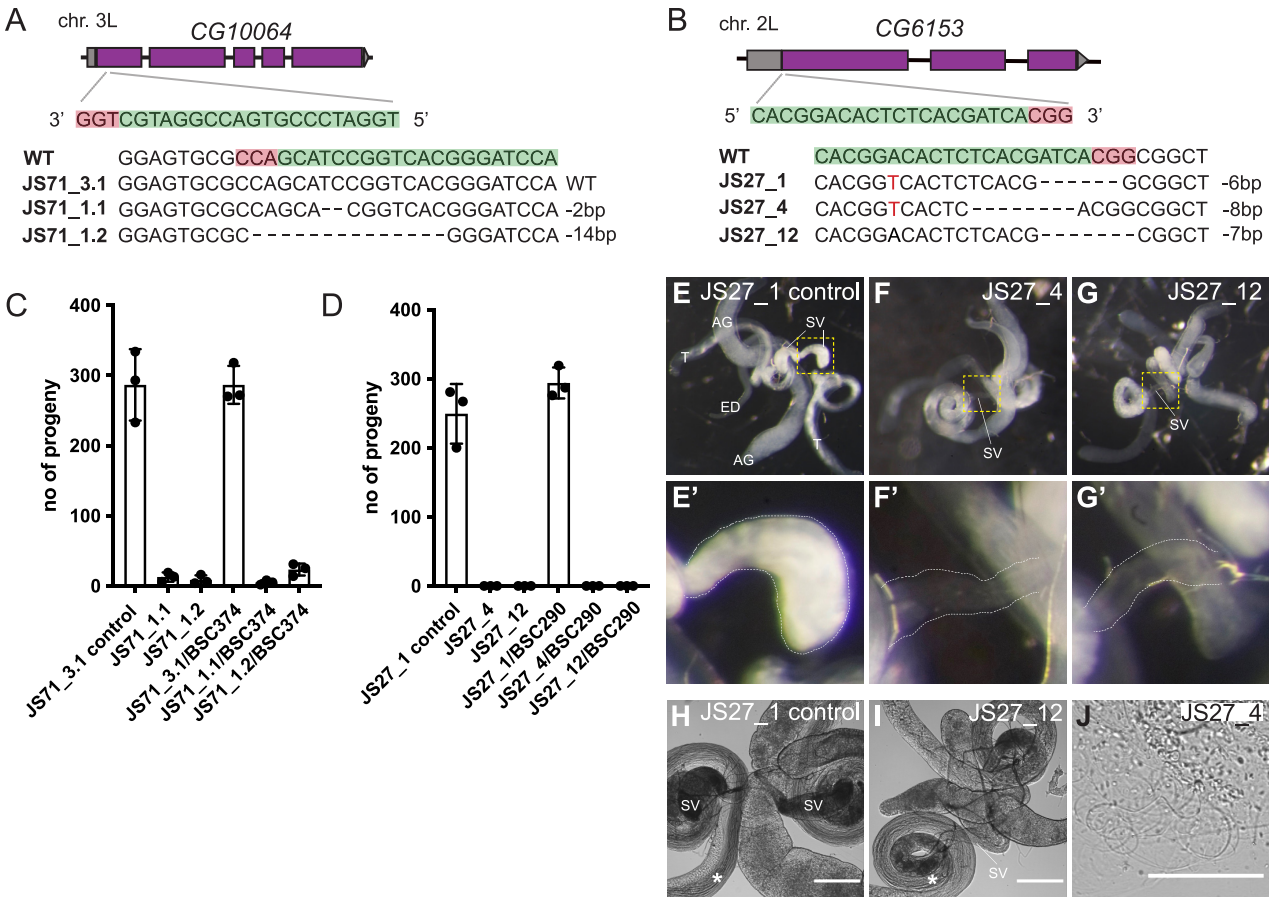



**Fig 6. Validation of RNAi male sterility phenotypes using CRISPR/Cas9 gene disruption.** (A, B) Schematics of the genomic locus of candidate genes, position of CRISPR target sites and mutant alleles analysed. (C, D) Assessment of male fertility of mutants (homozygous and over a deficiency). The graphs show mean values +/− SD of the number of progeny produced by mutant males. Three crosses with 5 wild-type virgins and 3 mutant males were analysed for each genotype. Wild-type males or males carrying in-frame mutations were used as controls. Where possible, alleles covering both alternative reading frames were analysed. (E–G) Widefield fluorescent micrographs of male reproductive systems of control and JS27/CG6153 mutants expressing Don Juan-GFP to label sperm. Mutants exhibit empty seminal vesicles, (E'-G') show zoomed regions of seminal vesicles from E–G (yellow dashed squares). (H–J) Widefield phase micrographs of reproductive systems of control and mutant males. Sperm are produced in both (asterisks), suggesting that sperm are made in the mutant but does not survive. Note that some mutant sperm gets into the ejaculatory duct (J). AG, accessory gland; ED, ejaculatory duct; SV, seminal vesicle; T, testis. Scale bars, 200 μm (H, I), 100 μm (J). The data underlying the graphs shown in the figure can be found in S1 Data. RNAi, RNA interference.

https://doi.org/10.1371/journal.pbio.3002222.g006

Notch signalling in embryos, although its role remains unclear [92]. The third related gene, *CG10795*, is also of unknown function, so we knocked it out with CRISPR/Cas-9 and discovered that it too showed phenotypes indicative of a severe defect in Notch signalling (Fig 7H–7L). Thus, all 3 proteins are required for a cellular process essential for embryonic Notch function, and recently, a similar conclusion was independently made by others [93]. All 3 human TM2D proteins were hits in a recent genome-wide screen for defects in endosomal function [94], and endosomes play a critical role in Notch signalling. Further work will be required to determine the precise role of these proteins, and how it relates to wing growth, but their likely role in endosomal function, combined with the existence of related TM2 domain proteins in bacteria and archaea, suggest fundamental roles in cell function rather than an exclusive role in Notch signalling.

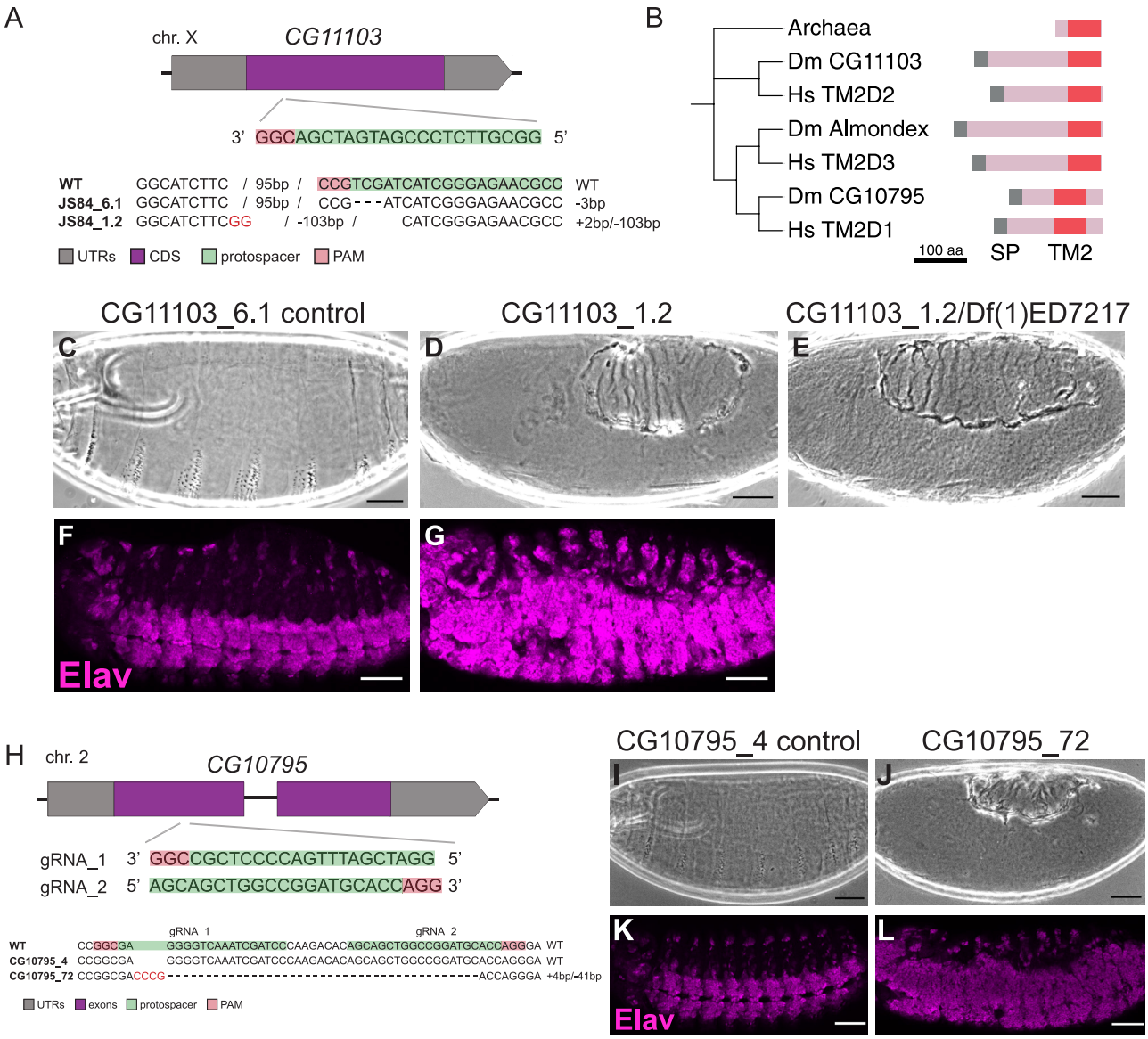



**Fig 7. Investigation of wing growth hit *CG11103* using CRISPR/Cas9 gene disruption.** (A) Schematic of the genomic locus of candidate *CG11103*, position of the CRISPR target site and the mutant allele analysed. Flies carrying an in-frame mutation were used as control. (B) Gene tree for TM2 domain proteins in humans and *Drosophila*, with an archaeal TM2 protein as an outlier. Tree built using sequence of TM2 domains alone using T-Coffee. A fourth TM2 domain protein is present in *Drosophila* and humans (Wurst/DNAJC22) which has additional TMDs and a DNAJ domain and appears to play a role in clathrin-mediated endocytosis [116]. (C–E) Cuticle phenotypes of embryos laid by control females and mutant females (homozygous or over a deficiency). (F, G) Micrographs of embryos laid by control females and homozygous mutant females stained against the pan-neuronal marker Elav. Scale bars: 50 μm. (H) Schematic of the genomic locus of *CG10795*, position of CRISPR target sites and the alleles analysed. Flies without an indel were used as control (*CG10795_4*). (I, J) Cuticle phenotypes of embryos laid by control or mutant females. (K, L) Micrographs of embryos laid by control or mutant females stained for the pan-neuronal marker Elav. Scale bars: 50 μm.

https://doi.org/10.1371/journal.pbio.3002222.g007

Taken together, this genetic validation data confirms that the RNAi screening approach, despite its known caveats, has given accurate phenotypic information for at least a substantial subset of the hits from our RNAi screens of the unknome set of genes.

## Discussion

The totality of scientific knowledge represents the summed activity of numerous individual research groups, each focusing on specific questions whose selection is influenced by many factors, some scientific and some more socially determined [7]. The latter set of factors includes issues like a preference for the relative safety, sociability, and kudos available when working in well-established fields, but is also strongly influenced by funding mechanisms. These usually aim to address societal needs but are subject to subjective assessment, historical precedent, and political pressures. In particular, the need to justify proposed research with reference to an established body of work, and preliminary data, may restrict investigation into truly unknown areas. Putting it more positively, there is potential for scientific progress to be accelerated by identifying situations where questions are being inadvertently and unjustifiably neglected. To quote James Clerk Maxwell "Thoroughly conscious ignorance is the prelude to every real advance in science." We have thus directly addressed here an area of long-standing concern: that biological research largely ignores less well known, but potentially important, genes [2,4,6,7]. Our results provide further evidence that this concern is well founded.

Our approach has been to develop an Unknome database. This has confirmed previous observations that poorly understood genes are relatively neglected; we also find that this problem is persisting even though there has been some progress in assigning functions to some of these genes. Recent developments in exome sequencing have allowed the identification of novel components of pathways whose genes give a well-defined set of disease symptoms, as has been seen with the cilia proteins identified from patients with ciliopathies [42,95]. In addition, the advent of the CRISPR/Cas9 system has enabled screens that cover whole genomes [17,96]. However, such screens are typically performed in cultured cells and hence cover only a subset of biological processes, and can also miss genes that have closely related, and thus functionally redundant, paralogs [97].

We used the Unknome database to select 260 genes that appeared both highly conserved and particularly poorly understood, and then applied functional assays in whole animals that would be impractical at genome-wide scale. Using 7 assays, designed to interrogate defects in a broad range of biological functions, we found phenotypes for 59 genes, in addition to the 62 genes that appear to be essential for viability (S2 Table and S4A and S4B Fig). Our approach relied on RNAi, but when 7 of the hits (corresponding to 6 genes) were retested with CRISPR/Cas9 gene disruption, we could validate 4. This is also a reminder that studies in model organisms such as *Drosophila* still have the scope to provide insight into unstudied human genes. The use of RNAi to knockdown candidate genes is powerful in this context because it allows for tissue-specific knockdown moreover, the likely incomplete loss of function achieved by RNAi can allow essential genes to reveal otherwise hidden hypomorphic phenotypes. Conversely, we note that as CRISPR approaches become ever more streamlined and sophisticated, future exploitation of the Unknome database can realistically use CRISPR technology to investigate functions of unknown genes.

An important primary conclusion of our work is that these uncharacterised genes have not deserved their neglect, a conclusion strengthened by a variety of other studies published during the protracted course of our studies, again revealing important functions for unknown genes. Again, this highlights the gradual shrinking, albeit slowly, of the unknome. Perhaps, most significantly, our database provides a powerful, versatile, and efficient platform to identify and select important genes of unknown function for analysis, thereby accelerating the closure of the gap in biological knowledge that the unknome represents. In practical terms, the Unknome database provides a resource for researchers who wish to exploit the opportunities associated

with unstudied areas of biology. Such endeavours will of course carry some risk as the outcome will be uncertain, and indeed, there is evidence that junior scientists are less likely to become principal investigators if they work on genes that have received little previous attention [7]. One approach may be collaborative efforts between labs to share resources and risk, and indeed, such an approach has recently been suggested by a consortium of proteomics groups [98].

Thinking about how to evaluate ignorance of gene function guided our bioinformatic approach to selecting of a set of genes small enough for complex phenotypic screening in a whole animal. At a broader level, we believe that acknowledging and evaluating ignorance is an important factor in decisions about the relative priority given to addressing the remaining fundamental questions in biology, versus translating and exploiting what we already know. However, ignorance can only have value if it can be meaningfully measured. Developing the Unknome database highlighted a couple of issues that affect our assessment of the state of knowledge of gene function. First, our approach relied on identifying orthologs from major organisms used for biological research. Although current methods for ortholog identification work well, there is still scope for improvement [24,25].

Secondly, our approach relied on the comprehensive and systematic annotation of gene function by the Gene Ontology (GO) Consortium [21,22]. Thus, another issue that arises from our work is that the current rapid rate of genome sequencing has required that most annotation is now automated rather than manual. This has led to the development of powerful methods to add functional annotation based on similarities to genes from other species [99]. However, such methods aim to cumulatively add annotation rather than remove disproven conclusions or address contradictions, which requires time-consuming manual curation. Moreover, increasing numbers of functional annotations are based on phenotypes from high-throughput screens for genetic phenotypes or protein–protein interactions, both of which are prone to generating false positives [100]. Thus, genes inevitably accrete annotations over time, some of which may be wrong, contradictory, or superficial but have little prospect of being corrected in the foreseeable future. As a result, the admirable aim of adding new gene annotation carries the risk of inadvertently obscuring our understanding of what is genuinely unknown.

An illustration of this problem is the gene CG9536 (TMEM115 in humans). This protein has been annotated as having endopeptidase activity based on distant sequence similarity to the rhomboid family of intramembrane proteases. However, CG9536, and its relatives in other species, lack the conserved residues that form the active site in rhomboids, and thus the only thing that can be currently concluded about the function of CG9536 is that it is almost certainly not a protease [101]. A more extreme case is htt, the *Drosophila* ortholog of huntingtin. This was not in the unknome test set because the extensive study of the role of huntingtin in human disease has led to many preliminary suggestions of function that have resulted in annotations linked to transcription, transport, autophagy, mitochondrial function, etc., and yet, the current consensus is that huntingtin's precise cellular role remains uncertain [102].

In conclusion, we find that accurately evaluating ignorance about gene function provides a valuable resource for guiding biological studies and may even be important for determining strategies to efficiently fund science. We have developed an approach to tackle directly the huge but under-discussed issue of the large number of well-conserved genes that have no reliably known function, despite the likelihood that they participate in major and even possibly completely new areas of biological function. We hope that our work will inspire others to define and characterise further the unknome and also to seek to ensure that gene annotation has the support and technology to preserve and recognise true ignorance.

## Materials and methods

### Construction of the Unknome database

The protein sequence data that we considered corresponds to the reference UniProt Proteomes [https://www.uniprot.org/proteomes/] used by the latest PANTHER database and includes human and 11 model organism species: *A. thaliana*, *C. elegans*, *D. rerio*. *D. discoideum*, *D. melanogaster*, *E. coli* (K12), *G. gallus*, *M. musculus*, *R. norvegicus*, *S. cerevisiae*, and *S. pombe* [26,103].

The Unknome database aggregates relevant information from the listed sources and provides a default knownness score for each protein and protein family (cluster) and can be recompiled in a few hours. Here, PANTHER provides the protein family information, via a group of UniProt IDs, that can be combined with selected information from UniProt entries, including protein sequence, GO terms, PubMed citations, species, gene name(s), and cross-references to species-specific databases.

The GO terms present in each UniProt entry were automatically provided by the Gene Ontology Annotation (GOA) database [https://www.ebi.ac.uk/GOA], based on GO release 2022-09-19 [22]. Evidence terms from the OBO Foundry are employed by GO [104], and in the Unknome database, they were weighted according to their evidence codes using the following default values: EXP; 0.8, IDA; 0.8, IPI; 0.8, IMP; 0.8, IGI; 0.8, IEP; 0.8, ISS; 0.5, ISO; 0.5, ISA; 0.5, ISM; 0.5, IGC; 0.3, RCA; 0.6, TAS; 0.9, NAS; 0.6, IC; 1.0, ND; 0.0, IEA; 0.0, NR; 0.0, IRD; 0.0, IKR; 0.0, IBA; 0.5, IBD; 0.5 (see http://geneontology.org/docs/guide-go-evidence-codes/ for a full description). After weighting, they were summed to generate a knownness score for each protein. The knownness score for the family defined by PANTHER is the maximum score among all the protein members present in the human and model organism list.

All protein GO terms linked in the database were dated according to when they were first linked with the UniProt entry, so as to be able to track the historical change of knownness. Though this information is not directly accessible within UniProt entries, the GOA database makes this information available via GAF format files at ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/. Note that this information only covers current entries and so annotations made in the past that were subsequently removed are not included in analyses of the change in knownness.

The Unknome is presented with a web interface at the URL http://unknome.org, with the entire database available to download as SQLite Version 3 files. This website is constructed using the Python module Django and provides views on the underlying database with easy filtering by knownness. In particular, the site displays the change over time in knownness for each protein cluster and lists the GO terms associated with each member of the cluster, along with their dates. The web site also makes all data available for download, from individual protein sequences to the whole SQL database file.

### *Drosophila* genetics

Hairpin RNAi stocks for the Unknome set were from the KK library of the Vienna Drosophila Resource Centre (S1 Table). During the course of our studies, it was reported that the stocks in this library have the transgene in one of 2 sites in the genome (the annotated locus 40D or the non-annotated site 30B), and insertions at 40D can cause lethality when the guide RNA is expressed [32,33]. PCR analysis with the previously used diagnostic primers was applied to 360 of the 365 lines, with the 5 remaining lines being lethal when expressed and so not included in any of the functional screens. This PCR analysis revealed that 98 of the 360 lines have the transgene in the problematic 40D site, a frequency of 27%, comparable to the 23% (9/39) and 25%

(38/150) found previously. All but one of these 98 lines gave a lethal or semi-lethal phenotype when crossed to the ubiquitous da-GAL4 driver (S1 Table).

Expression of the RNAi hairpins was driven with either the ubiquitous driver da-GAL4 driver, or with tissue-specific drivers: en-GAL4 (wing), bam-GAL4-VP16 (male fertility), MTD-GAL4 (female fertility), and GMR-GAL4 (proteostasis in the eye). UAS-Dicer-2 was included in all cases except for the 2 fertility screens as this has been found to improve the efficiency of RNAi [105]. For the proteostasis screen, the driver line also contained UAS-Httex1--Q46-eGFP [59]. In the lethality screen, those crosses that produced no adult progeny were defined as "lethal," while those where the progeny reached the pharate stage but the majority could not hatch, and those that did failed to expand wings and did not survive, were "semi-lethal."

For validation using CRISPR/Cas9, the following fly stocks were used: nos-phiC3; attP40 (DBSC #25709), nos-phiC3;;attP2 (DBSC #25710), CFD2 [106], TH_attP2 [107], Df(1)ED7217 (DBSC #8952), Df(2R)BSC268 (DBSC #26501), Df(2L)BSC812 (DBSC #27383), Df(2L) BSC290 (BDSC #23675), Df(3L)BSC374 (BDSC #24398). Spermatids and sperm were labelled with Don Juan (dj)-GFP [108].

## Fertility

Fertility was monitored using competitive assays, in which 1 red-eyed fly expressing the RNAi and 1 white-eyed w1118 fly were placed with 4 w1118 flies of the opposite sex. For male fertility, the Bam-Gal4 driver was used in combination with Dicer, and for female fertility, MTD-Gal4 was used without Dicer. RNAi stocks for the controls were from the VDRC: vret (GD 34897) and Ref1 (KK 10447). The flies were allowed to mate for 7 days, transferring to fresh vials every 2 to 3 days. After 7 days, the parental generation was removed and all progeny that emerged from the vial were counted, with eye colour used to determine the parent of each. Flies from the RNAi parent were separated, imaged, and quantified using Fiji image analysis platform [109], with a custom macro (https://github.com/tjs23/unknome). Individual data for both males and females were used to calculate means and the variances errors for the graphical plot (S2 and S3 Data).

## Wing growth assay

The genes in the unknome set were knocked down in the posterior half of the wing by using an engrailed-GAL4 driver combined with UAS-dcr-2. For each cross, at least 10 independent wings were collected and mounted on a slide under a coverslip in 50% glycerol/PBST. Images obtained with a 5× objective were analysed using a Fiji macro to contrast the veins from the rest of the wing (https://github.com/tjs23/unknome), and then, the areas of specific inter-vein regions in the anterior and posterior halves were determined. Individual data for each stock was used to calculate means and the variances errors for the graphical plot (S2 and S3 Data).

## Proteostasis assay in the eye

To interrogate the handling of misfolded proteins, a GFP fusion to part of huntingtin with a polyglutamine repeat was expressed in eyes, and the number of GFP-positive aggregates determined [59]. UAS-Httex1-Q46-eGFP was expressed in the eye along with the RNAi using GMR-Gal4. One eye from at least 10 males per genotype was imaged after 18 days at 25˚C, using 3 males per independent cross. GFP-positive aggregates were quantified with Fiji using a custom macro that determined the area of the eye and then scored aggregates that were either smaller or larger than 50 pixels (https://github.com/tjs23/unknome). Individual data for each

stock was used to calculate means and the variances errors for the graphical plot (S2 and S3 Data).

## Survival under stress

To measure lifespan under stress, we developed an automated system for following viability over many days. Flies were placed in 96-well plates and photographed every hour with image analysis then used to identify when the flies stopped moving. To prepare the plates, nitrogen-free fly food was placed at the bottom of each well (8 g agar, 50 g glucose, and 5 g pectin per litre with 0.25% nipagin, antibiotics, and 4 ml/litre propionic acid as preservative). To assay oxidative stress, the same food was used with the addition of 7.5 mM paraquat. Adult male flies were subdued with $CO_2$ and single flies placed in each well of the 96 well, with the plate sitting on ice to prevent escape before the plate was full. The plate was then sealed with gas permeant film. To image the plates over time, they were placed on a circular rotating platform and moved under a camera to be imaged every hour, with 3 such platforms or wheels arranged in a stack. At least 200 adults were assayed for each genotype, and custom Python scripts used to align the images of each plate and then track the movement of the flies in each well (https://github.com/tjs23/unknome). Lifespan was defined as the time point after the last change in position of the fly in the well. Individual data for both starvation and ROS conditions was used to calculate median survival times and the variances errors for the graphical plot (S2 and S3 Data).

## iFly climbing assay

The climbing speed of flies was measured using the iFly tracking system in which a single camera and mirrors are used to follow the movement of flies in a vial [75,76]. The RNAi stocks for the unknome set were crossed to the ubiquitous daughterless-Gal4 driver, and progeny collected at 8 days and 22 days post eclosion. The Pink1 control RNAi stock was from the VDRC (KK 109614). To follow locomotion, 8 flies were placed in a vial that was tapped to collect them at the bottom, and then placed in the iFly apparatus for filming over 30 s, with this repeated 3 times. Locomotion velocities were then determined using the iFly tracking software [80]. Individual data from both 8 days and 22 days was used to calculate means and the variances errors for the graphical plot (S2 and S3 Data).

## Summary of statistical methods

The general approach we took is as follows, with full details provided as Supporting information (S1 Text). We first modelled the distributions of the experimental results relating to each of the phenotypes under consideration parametrically. We thus formalised the goal of identifying outlying genes as identifying outlying sets of parameters corresponding to genes for each of the different phenotypes. Our approach involved 3 steps. First, we performed a regression to obtain estimates of the parameters for genes and an estimate of their variance–covariance matrix while controlling for batch and other effects. This was important because variability across batches was substantial for several of the phenotypes considered. The particular regression model used for this batch correction depended on the dataset.

The next step involved determining an outlier region. To do this, we transformed the parameter estimates so they more closely resembled a sample from a normal distribution such that an elliptical outlier region was appropriate. This transformation was often simply chosen as the identity, but in certain cases logistic transformations were used, for example. To describe how this region was determined, it will be helpful to fix the phenotype and write $\mu_1, \ldots, \mu_J$ for the unknown transformed parameters for the genes, where $J$ is the total number of genes

under consideration for that phenotype. Furthermore, let us write $\hat{\mu}_1, \ldots, \hat{\mu}_J$ for the corresponding (transformed) estimated parameters. Note that the $\mu_j$ were two-dimensional in most examples.

We modelled the $\mu_j$ as samples from a mixture of a normal distribution and a distribution of outliers and aimed to estimate the mean and variance matrix of this normal distribution to give the center and shape of the outlier region. The mean was estimated using a robust mean estimator applied to $\hat{\mu}_1, \ldots, \hat{\mu}_J$, such that the outlying genes did not influence the estimate. Analogously, we also obtained a robust estimate of the variance of the $(\hat{\mu}_j)_{j=1}^J$ to better reflect the variance of the bulk of the $(\mu_j)_{j=1}^J$. We then employed a bootstrap approach [110] to adjust this variance estimate to account for the sampling variability of the $(\hat{\mu}_j)_{j=1}^J$: The raw robust variance would be an overestimate of the corresponding quantity for the true transformed parameters.

Given the final mean and variance estimates, we took our outlier region to be the complement of the elliptical contour of a normal density with this mean and variance with a size such that the probability of falling outside the region was either 0.05 or 0.1, depending on the dataset. Note that in the cases where the parameters $\mu_j$ were one-dimensional, the ellipse was simply an interval. Finally, we performed a bootstrap hypothesis test for each gene $j$ with the null hypothesis being that $\mu_j$ falls within the outlier ellipse. We thus obtained $p$-values for each gene quantifying the evidence that it is an outlier according to the data. Note that this measure incorporates how outlying $\hat{\mu}_j$ is, but importantly it also takes into account the fact that $\hat{\mu}_j$ is a noisy estimate of the true $\mu_j$. These $p$-values were then corrected for multiple testing using the Benjamini–Hochberg procedure [111].

## CRISPR/Cas9-mediated knock-out

CRISPR target sites were chosen using the CRISPR Optimal Target Finder (http://targetfinder.flycrispr.neuro.brown.edu/). pCFD3 was used for BbsI-dependent gRNA cloning (http://www.crisprflydesign.org/) [106]. gRNA transgenics were generated for all candidate genes using BDSC stocks #25709 or #25710, depending on the chromosomal location of the target gene. To generate indels, transgenic gRNA lines were crossed to either CFD2 or TH_attP2. DNA microinjections were performed by the University of Cambridge Department of Genetics Fly Facility. For generation of *CG10795* mutants, gRNAs were cloned into pCFD3, and plasmids injected into CFD2 embryos. Stable stocks were generated to recover indels for all candidate genes. For genotyping, single males were collected and the genomic DNA was isolated using microLYSIS-Plus (Clent Life Science). Diagnostic PCRs followed by sequencing identified indels. Antibodies were not available to check protein levels, and so for those genes where we did not observe a phenotype, it is formally possible that residual or truncated protein was to blame.

## Fertility assays on CRISPR/Cas9 mutants

To check male fertility, crosses with 5 Oregon R wild-type virgins and 3 mutant males were set up for each genotype. Crosses were kept at 25˚C and knocked over twice. The total number of offspring was counted for all crosses and the mean +/− SD was plotted for each genotype. Deficiencies uncovering the candidate genes were used to check for potential off-target effects. To check female fertility, 3 crosses with 5 mutant virgins and 3 Oregon R wild-type males were set up for each genotype and processed in the same was as for male fertility. A deficiency uncovering *CG8237* was used to check for potential off-target effects.

### Analysis of *CG11103* and *CG10795* embryonic phenotypes

Overnight egg collections (at 25˚C) from *CG11103* and *CG10795* mutant females and males were kept at 25˚C for 48 h. Dead embryos were dechorionated and mounted in Hoyer's medium. Slides were kept at 65˚C for at least 24 h and widefield images obtained with a Zeiss Axioplan microscope. For examination of Elav expression, overnight egg collections from *CG11103* and *CG10795* mutant females and males were dechorionated with bleach and fixed using 4% formaldehyde. Embryos were devitellinised using n-Heptane/Methanol. Embryos were washed in PBS/0.1% Tween20 and blocked in PBS/0.1% Tween20 plus 5% BSA. Mouse anti-Elav (1/20; DSHB) were added over night at 4˚C, and then, embryos washed in PBS/0.1% Tween20. Donkey anti-mouse Alexa 488 (Fisher Scientific) was added and left for 2 h at RT. Embryos washed in PBS/0.1% Tween20 and mounted in Vectashield containing DAPI (Vector Laboratories) and imaged on a Zeiss LSM 710 confocal.

### Analysis of male seminal vesicles in *CG6153* mutants

Testes from 3 to 5 days old adult males were dissected in PBS and then either directly transferred onto a slide with Schneider's medium to take live images using a Zeiss 710 confocal microscope or fixed in 4% paraformaldehyde for 30 min at RT. PFA was then removed and the testes washed in PBT 0.1% Tween 20. Images were taken on a Zeiss stereomicroscope and a Nikon digital camera.

## Supporting information

**S1 Fig. Features of the Unknome database.** (A) Illustration of the interface in the Unknome database that can be used to weight GO annotations depending on the type of evidence. The settings shown are the default weightings that were used to generate an unknome gene set. (B) Clusters in the unknome that contain at least one human protein ranked by knownness, showing the distribution of proteins that are defined by Pfam as being in an uncharacterised protein family (UPF) or containing a domain of unknown function (DUF). The data underlying this graph can be found in S1 Data.
(EPS)

**S2 Fig. Trends in knownness.** (A) Change in the distribution of knownness of the 13,421 clusters that contain at least 1 protein from humans or the 11 model organisms. (B) Number of Gene Reference into Function (NCBI GeneRIF) annotations added per year since 2010 to the human genes in each of the 7,515 clusters that contain at least 1 human gene, ranked into deciles based on knownness in 2010. The best-known clusters in 2010 have received the most annotation in subsequent years. (C) Mean number of GO terms added to human-containing clusters per year for clusters ranked in deciles of knownness. The number of Process and Function GO terms added to all the genes in a cluster was summed and a mean determined for each year for all clusters in that centile. (D) Conservation in model organisms of human proteins in clusters as ranked in intervals of current knownness. (E) Mean number of species in each human-containing cluster as ranked in intervals of current knownness. Species are those in PANTHER 17.0, and better-known clusters tend to be present in a larger number of species. The data underlying the graphs shown in the figure can be found in S1 Data.
(EPS)

**S3 Fig. Testing of RNAi sterility hits using CRISPR/Cas9 gene disruption.** (A) Schematics of the genomic locus of candidate JS353/CG16890, position of CRISPR target sites and mutant alleles analysed. (B) Assessment of male fertility of CRISPR mutants in JS353/CG16890

(homozygous and over a deficiency). The graphs show mean values +/− SD of the number of progeny produced by mutant males. Three crosses with 5 WT virgins and 3 mutant males were analysed for each genotype. WT males or males carrying in-frame mutations were used as controls. Alleles covering both alternative reading frames were analysed. (C) Schematic of the genomic locus of candidate JS40/CG8237, position of the CRISPR target site and the mutant allele analysed. (D) Assessment of female fertility of mutants (homozygous and over a deficiency). The graph shows mean values +/− SD of the number of progeny produced by mutant females. Three crosses with 5 mutant virgins and 3 WT males were analysed. WT males and males carrying an in-frame mutation were used as controls. The data underlying the graphs shown in the figure can be found in S1 Data.
(EPS)

**S4 Fig. Graphical summary of the phenotypic screens.** (A) All genes that were analysed in the 7 phenotypic RNAi screens with those showing a phenotype in a screen indicated in red (see also S2 Table). For each screen, a few genes were omitted due to technical issues such as insufficient numbers of a particular cross being obtainable, or genes were analysed before they were found to be lethal and hence omitted from subsequent screens, and these are shown as blanks. The degree of conservation between each *Drosophila* protein and its human ortholog is indicated by the area of the circle shown. (B) Degree of amino conservation between the *Drosophila* proteins in the unknome set and their human orthologs, with the set that gave phenotypes (S2 Table), compared to those that did not. When there was more than 1 human ortholog in the cluster, the most closely related one was used. Relatedness calculated using the BLOSUM62 matrix. The data underlying the plot and the graph shown in the figure can be found in S1 Data.
(EPS)

**S1 Text. Supplemental text describing the statistical methods in depth.**
(PDF)

**S1 Table. List of *Drosophila* genes used for the unknome screen and the corresponding RNAi stocks.**
(XLSX)

**S2 Table. Genes whose knockdown gave significant effects in the functional screens.**
(XLSX)

**S1 Data. The data underlying the graphs and plots shown in Figs 1F, 2A, 2B, 4E, 6C and 6D, S1B, S2A–S2E, S3B, S3D and S4A–S4B Figs.**
(XLSX)

**S2 Data. Mean and variances from screens: Statistical analysis of batches assayed for each genotype in the functional screens, as used for plots in Figs 3A, 3D, 4B, 4F and 5C.**
(XLSX)

**S3 Data. Data points from screens: Data for individual flies from the batches assayed for each genotype in the functional screens as used to generate S2 Data.**
(XLSX)

**S1 Video. Representative time-lapse movie of a lifespan assay of flies in a 96-well plate.** Frames captured every hour and played at 30 frames/second.
(MP4)

## Acknowledgments

We thank Damian Crowther for loan of the iFly tracking system, Sara Imarisio for advice on proteostasis assays, Tobias Klöpper for help with gene selection for screens, the LMB workshops for help with the system for lifespan measurements, Anna Parish for fly stock maintenance, and Manu Hegde for comments on the manuscript.

## Author Contributions

**Conceptualization:** Matthew Freeman, Sean Munro.

**Formal analysis:** Tim J. Stevens, Rajen D. Shah, Sean Munro.

**Funding acquisition:** Matthew Freeman, Sean Munro.

**Investigation:** João J. Rocha, Satish Arcot Jayaram, Nadine Muschalik, Sahar Emran, Cristina Robles.

**Methodology:** João J. Rocha, Satish Arcot Jayaram, Rajen D. Shah.

**Project administration:** João J. Rocha, Matthew Freeman.

**Software:** João J. Rocha, Tim J. Stevens.

**Supervision:** Matthew Freeman, Sean Munro.

**Validation:** Nadine Muschalik.

**Visualization:** Tim J. Stevens.

**Writing – original draft:** Sean Munro.

**Writing – review & editing:** Matthew Freeman, Sean Munro.

## References

1. Adhikari S, Nice EC, Deutsch EW, Lane L, Omenn GS, Pennington SR, et al. A high-stringency blueprint of the human proteome. Nat Commun. 2020; 11:5301. https://doi.org/10.1038/s41467-020-19045-9 PMID: 33067450
2. Sinha S, Eisenhaber B, Jensen LJ, Kalbuaji B, Eisenhaber F. Darkness in the human gene and protein function space: widely modest or absent illumination by the life science literature and the trend for fewer protein function discoveries since 2000. Proteomics. 2018; 18:e1800093. https://doi.org/10.1002/pmic.201800093 PMID: 30265449
3. Wood V, Lock A, Harris MA, Rutherford K, Bähler J, Oliver SG. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? Open Biol. 2019; 9:180241. https://doi.org/10.1098/rsob.180241 PMID: 30938578
4. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. Nature. 2011; 470:163–165. https://doi.org/10.1038/470163a PMID: 21307913
5. Peña-Castillo L, Hughes TR. Why are there still over 1000 uncharacterized yeast genes? Genetics. 2007; 176:7–14. https://doi.org/10.1534/genetics.107.074468 PMID: 17435240
6. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, et al. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov. 2018; 17:317–332. https://doi.org/10.1038/nrd.2018.14 PMID: 29472638
7. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol. 2018; 16:e2006643. https://doi.org/10.1371/journal.pbio.2006643 PMID: 30226837
8. Firestein S. Ignorance: How It Drives Science. Oxford University Press; 2012.
9. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. Sci Rep. 2018;1–7. https://doi.org/10.1038/s41598-018-19333-x PMID: 29358745
10. Muñoz-Fuentes V, Cacheiro P, Meehan TF, Aguilar-Pimentel JA, Brown SDM, Flenniken AM, et al. The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian

genome that informs conservation. Conserv Genet Print. 2018; 19:995–1005. https://doi.org/10.1007/s10592-018-1072-9 PMID: 30100824

11. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Sci N Y NY. 2015; 347:1260419. https://doi.org/10.1126/science.1260419 PMID: 25613900

12. Rodgers G, Austin C, Anderson J, Pawlyk A, Colvis C, Margolis R, et al. Glimmers in illuminating the druggable genome. Nat Rev Drug Discov. 2018; 17:301–302. https://doi.org/10.1038/nrd.2017.252 PMID: 29348682

13. Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Res. 2017; 45:11495–11514. https://doi.org/10.1093/nar/gkx937 PMID: 29059321

14. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016; 17:184. https://doi.org/10.1186/s13059-016-1037-6 PMID: 27604469

15. Perdigão N, Rosa A. Dark proteome database: studies on dark proteins. High-Throughput. 2019; 8. https://doi.org/10.3390/ht8020008 PMID: 30934744

16. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021; 596:590–596. https://doi.org/10.1038/s41586-021-03828-1 PMID: 34293799

17. Wainberg M, Kamber RA, Balsubramani A, Meyers RM, Sinnott-Armstrong N, Hornburg D, et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat Genet. 2021; 53:638–649. https://doi.org/10.1038/s41588-021-00840-z PMID: 33859415

18. Duek P, Gateau A, Bairoch A, Lane L. Exploring the uncharacterized human proteome using neXtProt. J Proteome Res. 2018; 17:4211–4226. https://doi.org/10.1021/acs.jproteome.8b00537 PMID: 30191714

19. Nguyen D-T, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, et al. Pharos: Collating protein information to shed light on the druggable genome. Nucleic Acids Res. 2017; 45:D995–D1002. https://doi.org/10.1093/nar/gkw1072 PMID: 27903890

20. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database J Biol Databases Curation. 2016; 2016. https://doi.org/10.1093/database/baw100 PMID: 27374120

21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. https://doi.org/10.1038/75556 PMID: 10802651

22. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021; 49:D325–D334. https://doi.org/10.1093/nar/gkaa1113 PMID: 33290552

23. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. 2011;Chapter 6: Unit 6.12.1–19. https://doi.org/10.1002/0471250953.bi0612s35 PMID: 21901743

24. Wang Y, Yang S, Zhao J, Du W, Liang Y, Wang C, et al. Using Machine Learning to Measure Relatedness Between Genes: A Multi-Features Model. Sci Rep. 2019; 9:4192. https://doi.org/10.1038/s41598-019-40780-7 PMID: 30862804

25. Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldón T, Huerta-Cepas J, et al. Advances and Applications in the Quest for Orthologs. Mol Biol Evol. 2019:2157–2164. https://doi.org/10.1093/molbev/msz150 PMID: 31241141

26. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. 2022; 31:8–22. https://doi.org/10.1002/pro.4218 PMID: 34717010

27. Pfeiffer T, Hoffmann R. Temporal patterns of genes in scientific publications. Proc Natl Acad Sci U S A. 2007; 104:12052–12056. https://doi.org/10.1073/pnas.0701315104 PMID: 17620606

28. Holland LZ, Ocampo Daza D. A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution? Genome Biol. 2018; 19:209–4. https://doi.org/10.1186/s13059-018-1592-0 PMID: 30486862

29. Homem CCF, Steinmann V, Burkard TR, Jais A, Esterbauer H, Knoblich JA. Ecdysone and mediator change energy metabolism to terminate proliferation in *Drosophila* neural stem cells. Cell. 2014; 158:874–888. https://doi.org/10.1016/j.cell.2014.06.024 PMID: 25126791

30. Mummery-Widmer JL, Yamazaki M, Stoeger T, Novatchkova M, Bhalerao S, Chen D, et al. Genome-wide analysis of Notch signalling in *Drosophila* by transgenic RNAi. Nature. 2009; 458:987–992. https://doi.org/10.1038/nature07936 PMID: 19363474

31. Heigwer F, Port F, Boutros M. RNA Interference (RNAi) Screening in *Drosophila*. Genetics. 2018; 208:853–874. https://doi.org/10.1534/genetics.117.300077 PMID: 29487145

32. Green EW, Fedele G, Giorgini F, Kyriacou CP. A *Drosophila* RNAi collection is subject to dominant phenotypic effects. Nat Methods. 2014; 11:222–223. https://doi.org/10.1038/nmeth.2856 PMID: 24577271

33. Vissers JHA, Manning SA, Kulkarni A, Harvey KF. A *Drosophila* RNAi library modulates Hippo pathway-dependent tissue growth. Nat Commun. 2016; 7:10368. https://doi.org/10.1038/ncomms10368 PMID: 26758424

34. Czech B, Preall JB, McGinn J, Hannon GJ. A transcriptome-wide RNAi screen in the *Drosophila* ovary reveals factors of the germline piRNA pathway. Mol Cell. 2013; 50:749–761. https://doi.org/10.1016/j.molcel.2013.04.007 PMID: 23665227

35. Viswanatha R, Li Z, Hu Y, Perrimon N. Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in *Drosophila* cells. eLife. 2018; 7:705. https://doi.org/10.7554/eLife.36333 PMID: 30051818

36. Nishimura T, Fakim H, Brandmann T, Youn J-Y, Gingras A-C, Jinek M, et al. Human MARF1 is an endoribonuclease that interacts with the DCP1:2 decapping complex and degrades target mRNAs. Nucleic Acids Res. 2018; 46:12008–12021. https://doi.org/10.1093/nar/gky1011 PMID: 30364987

37. Yao Q, Cao G, Li M, Wu B, Zhang X, Zhang T, et al. Ribonuclease activity of MARF1 controls oocyte RNA homeostasis and genome integrity in mice. Proc Natl Acad Sci U S A. 2018; 115:11250–11255. https://doi.org/10.1073/pnas.1809744115 PMID: 30333187

38. Zhu L, Kandasamy SK, Liao SE, Fukunaga R. LOTUS domain protein MARF1 binds CCR4-NOT deadenylase complex to post-transcriptionally regulate gene expression in oocytes. Nat Commun. 2018; 9:4031. https://doi.org/10.1038/s41467-018-06404-w PMID: 30279526

39. Schulz J, Avci D, Queisser MA, Gutschmidt A, Dreher L-S, Fenech EJ, et al. Conserved cytoplasmic domains promote Hrd1 ubiquitin ligase complex formation for ER-associated degradation (ERAD). J Cell Sci. 2017; 130:3322–3335. https://doi.org/10.1242/jcs.206847 PMID: 28827405

40. Zhu B, Jiang L, Huang T, Zhao Y, Liu T, Zhong Y, et al. ER-associated degradation regulates Alzheimer's amyloid pathology and memory function by modulating γ-secretase activity. Nat Commun. 2017; 8:1472. https://doi.org/10.1038/s41467-017-01799-4 PMID: 29133892

41. Horani A, Ferkol TW. Advances in the genetics of primary ciliary dyskinesia: clinical implications. Chest. 2018; 154:645–652. https://doi.org/10.1016/j.chest.2018.05.007 PMID: 29800551

42. Legendre M, Zaragosi L-E, Mitchison HM. Motile cilia and airway disease. Semin Cell Dev Biol. 2021; 110:19–33. https://doi.org/10.1016/j.semcdb.2020.11.007 PMID: 33279404

43. Cheng W, Ip YT, Xu Z. Gudu, an Armadillo repeat-containing protein, is required for spermatogenesis in *Drosophila*. Gene. 2013; 531:294–300. https://doi.org/10.1016/j.gene.2013.08.080 PMID: 24055424

44. Diggle CP, Moore DJ, Mali G, zur Lage P, Ait-Lounis A, Schmidts M, et al. HEATR2 plays a conserved role in assembly of the ciliary motile apparatus. PLoS Genet. 2014; 10:e1004577. https://doi.org/10.1371/journal.pgen.1004577 PMID: 25232951

45. Coutton C, Vargas AS, Amiri-Yekta A, Kherraf Z-E, Ben Mustapha SF, Le Tanno P, et al. Mutations in CFAP43 and CFAP44 cause male infertility and flagellum defects in Trypanosoma and human. Nat Commun. 2018; 9:686. https://doi.org/10.1038/s41467-017-02792-7 PMID: 29449551

46. Ta-Shma A, Perles Z, Yaacov B, Werner M, Frumkin A, Rein AJJT, et al. A human laterality disorder associated with a homozygous WDR16 deletion. Eur J Hum Genet EJHG. 2015; 23:1262–1265. https://doi.org/10.1038/ejhg.2014.265 PMID: 25469542

47. Gui L, Song K, Tritschler D, Bower R, Yan S, Dai A, et al. Scaffold subunits support associated subunit assembly in the Chlamydomonas ciliary nexin-dynein regulatory complex. Proc Natl Acad Sci U S A. 2019; 116:23152–23162. https://doi.org/10.1073/pnas.1910960116 PMID: 31659045

48. Kravtsova-Ivantsiv Y, Shomer I, Cohen-Kaplan V, Snijder B, Superti-Furga G, Gonen H, et al. KPC1-mediated ubiquitination and proteasomal processing of NF-κB1 p105 to p50 restricts tumor growth. Cell. 2015; 161:333–347. https://doi.org/10.1016/j.cell.2015.03.001 PMID: 25860612

49. Li W, Liang J, Outeda P, Turner S, Wakimoto BT, Watnick T. A genetic screen in *Drosophila* reveals an unexpected role for the KIP1 ubiquitination-promoting complex in male fertility. PLoS Genet. 2020; 16:e1009217. https://doi.org/10.1371/journal.pgen.1009217 PMID: 33378371

50. Hahn I, Fuss B, Peters A, Werner T, Sieberg A, Gosejacob D, et al. The *Drosophila* Arf GEF Steppke controls MAPK activation in EGFR signaling. J Cell Sci. 2013; 126:2470–2479. https://doi.org/10.1242/jcs.120964 PMID: 23549788

51. Ibar C, Glavic A. *Drosophila* p115 is required for Cdk1 activation and G2/M cell cycle transition. Mech Dev. 2017; 144:191–200. https://doi.org/10.1016/j.mod.2017.04.001 PMID: 28396045

52. Böhni R, Riesgo-Escovar J, Oldham S, Brogiolo W, Stocker H, Andruss BF, et al. Autonomous control of cell and organ size by CHICO, a *Drosophila* homolog of vertebrate IRS1-4. Cell. 1999; 97:865–875.

53. Irvine KD, Harvey KF. Control of organ growth by patterning and hippo signaling in *Drosophila*. Cold Spring Harb Perspect Biol. 2015; 7. https://doi.org/10.1101/cshperspect.a019224 PMID: 26032720

54. Bar-Peled L, Chantranupong L, Cherniack AD, Chen WW, Ottina KA, Grabiner BC, et al. A Tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. Sci N Y NY. 2013; 340:1100–1106. https://doi.org/10.1126/science.1232044 PMID: 23723238

55. Wei Y, Reveal B, Cai W, Lilly MA. The GATOR1 Complex Regulates Metabolic Homeostasis and the Response to Nutrient Stress in *Drosophila melanogaster*. G3 Bethesda Md. 2016; 6:3859–3867. https://doi.org/10.1534/g3.116.035337 PMID: 27672113

56. Hjeij R, Onoufriadis A, Watson CM, Slagle CE, Klena NT, Dougherty GW, et al. CCDC151 mutations cause primary ciliary dyskinesia by disruption of the outer dynein arm docking complex formation. Am J Hum Genet. 2014; 95:257–274. https://doi.org/10.1016/j.ajhg.2014.08.005 PMID: 25192045

57. Michellod M-A, Randsholt NB. Implication of the *Drosophila* beta-amyloid peptide binding-like protein AMX in Notch signaling during early neurogenesis. Brain Res Bull. 2008; 75:305–309. https://doi.org/10.1016/j.brainresbull.2007.10.060 PMID: 18331889

58. Russo A. Understanding the mammalian TRAP complex function(s). Open Biol. 2020; 10:190244. https://doi.org/10.1098/rsob.190244 PMID: 32453970

59. Zhang S, Binari R, Zhou R, Perrimon N. A genomewide RNA interference screen for modifiers of aggregates formation by mutant Huntingtin in *Drosophila*. Genetics. 2010; 184:1165–1179. https://doi.org/10.1534/genetics.109.112516 PMID: 20100940

60. Eidhof I, Baets J, Kamsteeg E-J, Deconinck T, van Ninhuijs L, Martin J-J, et al. GDAP2 mutations implicate susceptibility to cellular stress in a new form of cerebellar ataxia. Brain. 2018; 141:2592–2604. https://doi.org/10.1093/brain/awy198 PMID: 30084953

61. Farhan SMK, Nixon KCJ, Everest M, Edwards TN, Long S, Segal D, et al. Identification of a novel synaptic protein, TMTC3, involved in periventricular nodular heterotopia with intellectual disability and epilepsy. Hum Mol Genet. 2017; 26:4278–4289. https://doi.org/10.1093/hmg/ddx316 PMID: 28973161

62. Li J, Akil O, Rouse SL, McLaughlin CW, Matthews IR, Lustig LR, et al. Deletion of Tmtc4 activates the unfolded protein response and causes postnatal hearing loss. J Clin Invest. 2018; 128:5150–5162. https://doi.org/10.1172/JCI97498 PMID: 30188326

63. Hamdan N, Kritsiligkou P, Grant CM. ER stress causes widespread protein aggregation and prion formation. J Cell Biol. 2017; 216:2295–2304. https://doi.org/10.1083/jcb.201612165 PMID: 28630146

64. Fujiwara T, Ye S, Castro-Gomes T, Winchell CG, Andrews NW, Voth DE, et al. PLEKHM1/DEF8/RAB7 complex regulates lysosome positioning and bone homeostasis. JCI Insight. 2016; 1:e86330. https://doi.org/10.1172/jci.insight.86330 PMID: 27777970

65. Gillingham AK, Sinka R, Torres IL, Lilley KS, Munro S. Toward a comprehensive map of the effectors of Rab GTPases. Dev Cell. 2014; 31:358–373. https://doi.org/10.1016/j.devcel.2014.10.007 PMID: 25453831

66. Pugh RJ, Slee JB, Farwell SLN, Li Y, Barthol T, Patton WA, et al. Transmembrane Protein 184A Is a Receptor Required for Vascular Smooth Muscle Cell Responses to Heparin. J Biol Chem. 2016; 291:5326–5341. https://doi.org/10.1074/jbc.M115.681122 PMID: 26769966

67. Ong YS, Tran THT, Gounko NV, Hong W. TMEM115 is an integral membrane protein of the Golgi complex involved in retrograde transport. J Cell Sci. 2014; 127:2825–2839. https://doi.org/10.1242/jcs.136754 PMID: 24806965

68. Takar M, Huang Y, Graham TR. The PQ-loop protein Any1 segregates Drs2 and Neo1 functions required for viability and plasma membrane phospholipid asymmetry. J Lipid Res. 2019;jlr.M093526. https://doi.org/10.1194/jlr.M093526 PMID: 30824614

69. Lee W-H, Higuchi H, Ikeda S, Macke EL, Takimoto T, Pattnaik BR, et al. Mouse Tmem135 mutation reveals a mechanism involving mitochondrial dynamics that leads to age-dependent retinal pathologies. eLife. 2016; 5:7618. https://doi.org/10.7554/eLife.19264 PMID: 27863209

70. Shibano T, Mamada H, Hakuno F, Takahashi S-I, Taira M. The Inner Nuclear Membrane Protein Nemp1 Is a New Type of RanGTP-Binding Protein in Eukaryotes. PLoS ONE. 2015; 10:e0127271. https://doi.org/10.1371/journal.pone.0127271 PMID: 25946333

71. Zhang K, Li Z, Jaiswal M, Bayat V, Xiong B, Sandoval H, et al. The C8ORF38 homologue Sicily is a cytosolic chaperone for a mitochondrial complex I subunit. J Cell Biol. 2013; 200:807–820. https://doi.org/10.1083/jcb.201208033 PMID: 23509070

72. Phillips JP, Campbell SD, Michaud D, Charbonneau M, Hilliker AJ. Null mutation of copper/zinc superoxide dismutase in *Drosophila* confers hypersensitivity to paraquat and reduced longevity. Proc Natl Acad Sci U S A. 1989; 86:2761–2765.

73. Rzezniczak TZ, Douglas LA, Watterson JH, Merritt TJS. Paraquat administration in *Drosophila* for use in metabolic studies of oxidative stress. Anal Biochem. 2011; 419:345–347. https://doi.org/10.1016/j.ab.2011.08.023 PMID: 21910964

74. Guan J-J, Zhang X-D, Sun W, Qi L, Wu J-C, Qin Z-H. DRAM1 regulates apoptosis through increasing protein levels and lysosomal localization of BAX. Cell Death Dis. 2015; 6:e1624. https://doi.org/10.1038/cddis.2014.546 PMID: 25633293

75. Secchi C, Carta M, Crescio C, Spano A, Arras M, Caocci G, et al. T cell tyrosine phosphorylation response to transient redox stress. Cell Signal. 2015; 27:777–788. https://doi.org/10.1016/j.cellsig.2014.12.014 PMID: 25572700

76. Srinivasan N, Gordon O, Ahrens S, Franz A, Deddouche S, Chakravarty P, et al. Actin is an evolutionarily-conserved damage-associated molecular pattern that signals tissue injury in *Drosophila melanogaster*. eLife. 2016; 5:72. https://doi.org/10.7554/eLife.19662 PMID: 27871362

77. Tsygankov AY. TULA-family proteins: Jacks of many trades and then some. J Cell Physiol. 2018; 234:274–288. https://doi.org/10.1002/jcp.26890 PMID: 30076707

78. Jana S, Hsieh AC, Gupta R. Reciprocal amplification of caspase-3 activity by nuclear export of a putative human RNA-modifying protein, PUS10 during TRAIL-induced apoptosis. Cell Death Dis. 2017; 8:e3093. https://doi.org/10.1038/cddis.2017.476 PMID: 28981101

79. Jahn TR, Kohlhoff KJ, Scott M, Tartaglia GG, Lomas DA, Dobson CM, et al. Detection of early locomotor abnormalities in a *Drosophila* model of Alzheimer's disease. J Neurosci Methods. 2011; 197:186–189. https://doi.org/10.1016/j.jneumeth.2011.01.026 PMID: 21315762

80. Kohlhoff KJ, Jahn TR, Lomas DA, Dobson CM, Crowther DC, Vendruscolo M. The iFly tracking system for an automated locomotor and behavioural analysis of *Drosophila melanogaster*. Integr Biol Quant Biosci Nano Macro. 2011; 3:755–760. https://doi.org/10.1039/c0ib00149j PMID: 21698336

81. McNally KE, Faulkner R, Steinberg F, Gallon M, Ghai R, Pim D, et al. Retriever is a multiprotein complex for retromer-independent endosomal cargo recycling. Nat Cell Biol. 2017; 19:1214–1225. https://doi.org/10.1038/ncb3610 PMID: 28892079

82. Voineagu I, Huang L, Winden K, Lazaro M, Haan E, Nelson J, et al. CCDC22: a novel candidate gene for syndromic X-linked intellectual disability. Mol Psychiatry. 2012; 17:4–7. https://doi.org/10.1038/mp.2011.95 PMID: 21826058

83. Matta JA, Gu S, Davini WB, Lord B, Siuda ER, Harrington AW, et al. NACHO mediates nicotinic acetylcholine receptor function throughout the brain. Cell Rep. 2017; 19:688–696. https://doi.org/10.1016/j.celrep.2017.04.008 PMID: 28445721

84. McNabb S, Greig S, Davis T. The alcohol dehydrogenase gene is nested in the outspread locus of *Drosophila melanogaster*. Genetics. 1996; 143:897–911.

85. Surks HK, Riddick N, Ohtani K-I. M-RIP targets myosin phosphatase to stress fibers to regulate myosin light chain phosphorylation in vascular smooth muscle cells. J Biol Chem. 2005; 280:42543–42551. https://doi.org/10.1074/jbc.M506863200 PMID: 16257966

86. Tapia Contreras C, Hoyer-Fender S. The WD40-protein CFAP52/WDR16 is a centrosome/basal body protein and localizes to the manchette and the flagellum in male germ cells. Sci Rep. 2020; 10:14240. https://doi.org/10.1038/s41598-020-71120-9 PMID: 32859975

87. Andersen KM, Madsen L, Prag S, Johnsen AH, Semple CA, Hendil KB, et al. Thioredoxin Txnl1/TRP32 is a redox-active cofactor of the 26 S proteasome. J Biol Chem. 2009; 284:15246–15254. https://doi.org/10.1074/jbc.M900016200 PMID: 19349277

88. Wiseman RL, Chin K-T, Haynes CM, Stanhill A, Xu C-F, Roguev A, et al. Thioredoxin-related Protein 32 is an arsenite-regulated Thiol Reductase of the proteasome 19 S particle. J Biol Chem. 2009; 284:15233–15245. https://doi.org/10.1074/jbc.M109.002121 PMID: 19349280

89. Kondo H, Matsumura T, Kaneko M, Inoue K, Kosako H, Ikawa M, et al. PITHD1 is a proteasome-interacting protein essential for male fertilization. J Biol Chem. 2020; 295:1658–1672. https://doi.org/10.1074/jbc.RA119.011144 PMID: 31915251

90. Lachén-Montes M, Mendizuri N, Ausín K, Pérez-Mediavilla A, Azkargorta M, Iloro I, et al. Smelling the Dark Proteome: Functional Characterization of PITH Domain-Containing Protein 1 (C1orf128) in Olfactory Metabolism. J Proteome Res. 2020; 19:4826–4843. https://doi.org/10.1021/acs.jproteome.0c00452 PMID: 33185454

91. Kajkowski EM, Lo CF, Ning X, Walker S, Sofia HJ, Wang W, et al. beta -Amyloid peptide-induced apoptosis regulated by a novel protein containing a g protein activation module. J Biol Chem. 2001; 276:18748–18756. https://doi.org/10.1074/jbc.M011161200 PMID: 11278849

92. Michellod M-A, Forquignon F, Santamaria P, Randsholt NB. Differential requirements for the neurogenic gene almondex during *Drosophila melanogaster* development. Genesis. 2003; 37:113–122. https://doi.org/10.1002/gene.10233 PMID: 14595834

93. Salazar JL, Yang SA, Lin YQ, Li-Kroeger D, Marcogliese PC, Deal SL, et al. TM2D genes regulate Notch signaling and neuronal function in *Drosophila*. PLoS Genet. 2021; 17:e1009962. https://doi.org/10.1371/journal.pgen.1009962 PMID: 34905536

94. Haney MS, Bohlen CJ, Morgens DW, Ousey JA, Barkal AA, Tsui CK, et al. Identification of phagocytosis regulators using magnetic genome-wide CRISPR screens. Nat Genet. 2018:1–16. https://doi.org/10.1038/s41588-018-0254-1 PMID: 30397336

95. Horani A, Ferkol TW, Dutcher SK, Brody SL. Genetics and biology of primary ciliary dyskinesia. Paediatr Respir Rev. 2016; 18:18–24. https://doi.org/10.1016/j.prrv.2015.09.001 PMID: 26476603

96. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a Cancer Dependency Map. Cell. 2017; 170:564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010 PMID: 28753430

97. De Kegel B, Ryan CJ. Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. PLoS Genet. 2019; 15:e1008466. https://doi.org/10.1371/journal.pgen.1008466 PMID: 31652272

98. Kustatscher G, Collins T, Gingras A-C, Guo T, Hermjakob H, Ideker T, et al. Understudied proteins: opportunities and challenges for functional proteomics. Nat Methods. 2022; 19:774–779. https://doi.org/10.1038/s41592-022-01454-x PMID: 35534633

99. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013; 10:221–227. https://doi.org/10.1038/nmeth.2340 PMID: 23353650

100. Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol. 2013; 9:e1003063. https://doi.org/10.1371/journal.pcbi.1003063 PMID: 23737737

101. Freeman M. The rhomboid-like superfamily: molecular mechanisms and biological roles. Annu Rev Cell Dev Biol. 2014; 30:235–254. https://doi.org/10.1146/annurev-cellbio-100913-012944 PMID: 25062361

102. Barron JC, Hurley EP, Parsons MP. Huntingtin and the Synapse. Front Cell Neurosci. 2021; 15:689332. https://doi.org/10.3389/fncel.2021.689332 PMID: 34211373

103. Consortium UniProt. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021; 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100 PMID: 33237286

104. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007; 25:1251–1255. https://doi.org/10.1038/nbt1346 PMID: 17989687

105. Dietzl G, Chen D, Schnorrer F, Su K-C, Barinova Y, Fellner M, et al. A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. Nature. 2007; 448:151–156. https://doi.org/10.1038/nature05954 PMID: 17625558

106. Port F, Chen H-M, Lee T, Bullock SL. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in *Drosophila*. Proc Natl Acad Sci U S A. 2014; 111:E2967–76. https://doi.org/10.1073/pnas.1405500111 PMID: 25002478

107. Port F, Muschalik N, Bullock SL. Systematic evaluation of *Drosophila* CRISPR tools reveals safe and robust alternatives to autonomous gene drives in basic research. G3 Bethesda Md. 2015; 5:1493–1502. https://doi.org/10.1534/g3.115.019083 PMID: 25999583

108. Santel A, Winhauer T, Blumer N, RenkawitzPohl R. The *Drosophila* don juan (dj) gene encodes a novel sperm specific protein component characterized by an unusual domain of a repetitive amino acid motif. Mech Dev. 1997; 64:19–30. https://doi.org/10.1016/s0925-4773(97)00031-2 PMID: 9232593

109. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012; 9:676–682. https://doi.org/10.1038/nmeth.2019 PMID: 22743772

110. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press; 1994.

111. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995; 57:289–300.

**112.** Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Valencia A, editor. Bioinformatics. 2020; 36:2628–2629. https://doi.org/10.1093/bioinformatics/btz931 PMID: 31882993

**113.** Zamparini AL, Davis MY, Malone CD, Vieira E, Zavadil J, Sachidanandam R, et al. Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in *Drosophila*. Development. 2011; 138:4039–4050. https://doi.org/10.1242/dev.069187 PMID: 21831924

**114.** Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, Mozden N, et al. The Berkeley *Drosophila* genome project gene disruption project: single P-element insertions mutating 25% of vital *Drosophila* genes. Genetics. 1999; 153:135–177. https://doi.org/10.1093/genetics/153.1.135 PMID: 10471706

**115.** Park J, Lee SB, Lee S, Kim Y, Song S, Kim S, et al. Mitochondrial dysfunction in *Drosophila* PINK1 mutants is complemented by parkin. Nature. 2006; 441:1157–1161. https://doi.org/10.1038/nature04788 PMID: 16672980

**116.** Behr M, Wingen C, Wolf C, Schuh R, Hoch M. Wurst is essential for airway clearance and respiratory-tube size control. Nat Cell Biol. 2007; 9:847–853. https://doi.org/10.1038/ncb1611 PMID: 17558392