Molecular Biology and Evolution, 2025, **42**, 1–22
https://doi.org/10.1093/molbev/msaf097
Advance access publication 25 April 2025
Discoveries

MBE OXFORD

Transposable Elements Drive Regulatory and Functional Innovation of F-box Genes

Miguel Vasconcelos Almeida [1,2,*], Zixin Li [3,7], Pedro Rebelo-Guiomar [1], Alexandra Dallaire [1,2,4], Lukáš Fiedler [1,2], Jonathan L. Price [1,2], Jan Sluka [5,6], Xiaodan Liu [1,2], Falk Butter [5,6], Christian Rödelsperger [*3], Eric A. Miska [1,2,*]

[1]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK
[2]The Gurdon Institute, University of Cambridge, Cambridge CB2 1QN, UK
[3]Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Tübingen 72076, Germany
[4]Comparative Fungal Biology, Royal Botanic Gardens Kew, Jodrell Laboratory, Richmond TW9 3DS, UK
[5]Institute of Molecular Biology (IMB), Quantitative Proteomics, Mainz 55128, Germany
[6]Institute of Molecular Virology and Cell Biology, Friedrich-Loeffler-Institute, Südufer, Greifswald 17493, Germany
[7]Present address: Department of Fish Ecology and Evolution, Swiss Federal Institute of Aquatic Science and Technology (EAWAG), Kastanienbaum 6047, Switzerland

***Corresponding authors:** E-mails: mdd34@cam.ac.uk; christian.roedelsperger@tuebingen.mpg.de; eam29@cam.ac.uk.
Associate editor: Grace Yuh Chwen Lee

Abstract

Protein domains of transposable elements (TEs) and viruses increase the protein diversity of host genomes by recombining with other protein domains. By screening 10 million eukaryotic proteins, we identified several domains that define multicopy gene families and frequently co-occur with TE/viral domains. Among these, a Tc1/Mariner transposase helix-turn-helix (HTH) domain was captured by F-box genes in the *Caenorhabditis* genus, creating a new class of F-box genes. For specific members of this class, like *fbxa-215*, we found that the HTH domain is required for diverse processes including germ granule localization, fertility, and thermotolerance. Furthermore, we provide evidence that Heat Shock Factor 1 (HSF-1) mediates the transcriptional integration of *fbxa-215* into the heat shock response by binding to Helitron TEs directly upstream of the *fbxa-215* locus. The interactome of HTH-bearing F-box factors suggests roles in post-translational regulation and proteostasis, consistent with established functions of F-box proteins. Based on AlphaFold2 multimer proteome-wide screens, we propose that the HTH domain may diversify the repertoire of protein substrates that F-box factors regulate post-translationally. We also describe an independent capture of a TE domain by F-box genes in zebrafish. In conclusion, we identify two independent TE domain captures by F-box genes in eukaryotes and provide insights into how these novel proteins are integrated within host gene regulatory networks.

Keywords: transposable elements, genome evolution, gene families, F-box genes, *Caenorhabditis elegans*

Introduction

In an influential analogy, Jacob (1977) argued that evolution does not work like an engineer, striving for perfection in their creations. Instead, he asserted evolution acts as tinkerer "who does not know exactly what he is going to produce but uses whatever he finds around him". In this sense, novelty is seldom created fully anew, de novo, but through the recombination of preexisting material. He argued that "to create is to recombine" (Jacob 1977). These principles can be applied to protein evolution, with novel proteins arising by recombination of preexisting modules. Several mechanisms can drive the recombination of protein modules, including exon shuffling, gene fusion, and transposition (Gilbert 1978; Patthy 1999; Babushok et al. 2007; Long et al. 2013).

Transposable elements (TEs) are mobile genetic elements detectable in most sequenced genomes, and often encode specialized protein machinery that is employed in their mobilization (Wells and Feschotte 2020). There are many distinct types of TEs, categorized according to their sequence features, proteins encoded, and replication mechanism (Wicker et al. 2007; Bourque et al. 2018; Wells and Feschotte 2020). In brief, TEs are commonly divided in two broad classes based on the mechanism of mobilization: class I TEs, or retrotransposons, transpose via an RNA intermediate, whereas class II TEs, or DNA transposons, employ a variety of mobilization mechanisms exclusively via DNA intermediates (Wicker et al. 2007; Bourque et al. 2018; Wells and Feschotte 2020).

TEs are recognized as a major force in the evolution of eukaryotic genomes, driving innovation in a variety of ways (Bourque et al. 2018; Wells and Feschotte 2020; Almeida et al. 2022; Fueyo et al. 2022). TEs often contain and disperse transcription factor binding sites across eukaryotic genomes (Bourque et al. 2018; Almeida et al. 2022; Fueyo et al. 2022). This leads to their frequent repurposing as promoters or enhancers of endogenous genes, with the potential to establish or rewire cis-regulatory networks. The proteins encoded by TEs are also a source of innovation. They can contribute to the generation of novel proteins by recombining with preexisting protein domains. For example, SETMAR proteins originated in primates via fusion of a Mariner DNA transposon with an SET histone methyltransferase (Robertson and Zumpano 1997; Cordaux et al. 2006). As another example

Received: December 18, 2024. **Revised:** February 27, 2025. **Accepted:** April 3, 2025

in primates, CSB-PGBD3 emerged from a PiggyBac DNA transposon fusion with the Cockayne syndrome group B (CSB) gene (Newman et al. 2008). Larger-scale computational surveys of available genomic, transcriptomic, and proteomic data have identified additional transcripts and proteins with co-occurring TE- and host-derived protein domains (Zdobnov et al. 2005; Cosby et al. 2021; Coronado-Zamora and González 2023; Oggenfuss et al. 2024). One study, focusing on host-transposase fusion genes in tetrapod evolution, reported a tendency of TE-derived DNA-binding domains to fuse to host domains associated with the regulation of gene expression (Cosby et al. 2021). This study described KRABINER, a host-transposase fusion gene that binds DNA and regulates gene expression, akin to SETMAR and CSB-PGBD3 (Cordaux et al. 2006; Bailey et al. 2012; Gray et al. 2012; Tellier and Chalmers 2019; Cosby et al. 2021).

The increasing availability of sequenced genomes, along with their protein-coding and repetitive element annotations, provides a massive publicly available resource that can be leveraged to identify protein innovations. In this study, we screened 10 million eukaryotic proteins to identify biologically relevant TE- and virus-derived novelties. We describe two phylogenetically independent captures of TE domains by F-box genes in animals, and characterize in detail one instance in nematodes. In this case, one domain derived from Tc1/Mariner TEs was captured by an F-box gene and created a novel F-box gene family. These genes require the TE domain for diverse functions, including thermotolerance, fertility, and germ granule localization. We further describe how Helitron TEs integrate one of these genes in a heat-stress-responsive pathway.

Results

Multicopy Protein Domains Recurrently Capture Protein Domains Associated With TEs and Viruses

We reasoned that protein domains that usually exist in multiple copies in genomes, and are typically associated with multigene families, may retain fusions with TE- or virus-derived protein domains more often. The relaxed selective pressure on these fusions may allow them to persist, even if the fusions are initially detrimental. This reasoning prompted our search for novel proteins with these domain architectures. We included protein domains associated with viruses in addition to TE domains, because long terminal repeat (LTR) retrotransposons are evolutionarily related to retroviruses, thus blurring the boundaries between these genetic elements (Eickbush and Malik 2007; Wells and Feschotte 2020).

To find instances of co-occurrence of TE- and virus-derived protein domains with multicopy domains, we searched ~10 million unique eukaryotic proteins from the UniProt database (The UniProt Consortium 2023) and found 12,803 unique proteins in the major eukaryotic clades with such domain architectures (Fig. 1a and supplementary table S1, Supplementary Material online). Of these, 9,471 have only one TE/viral domain, while 3,332 proteins display co-occurrence with multiple TE/viral domains, suggesting more complex fusions, possibly when several domains of one TE/virus are captured simultaneously (supplementary fig. S1a, Supplementary Material online). There is a strong association between the size of the protein domain family and the number of co-occurring TE/virus-derived domains (supplementary fig. S1b, Supplementary Material online). Reverse-transcriptase, integrase, and peptidase domains are highly represented, associating with most multicopy domains (Fig. 1b, c). In turn, protein kinase, leucine-rich repeats, ankyrin repeats, GPCR, and F-box domains were amongst the multicopy domains associated with most TE/viral domains (Fig. 1d). The majority of the co-occurring domain combinations are found in only a small number of taxa, suggesting that these domain fusions happened relatively recently in evolutionary time (supplementary fig. S1c, Supplementary Material online).

To validate our approach, we searched for known instances of TE- or virus-derived domain fusions with multicopy domains. The largest set of proteins with the same co-occurring domains, a BTB/POZ domain (IPR011333) with an integrase zinc-binding domain (IPR041588, Fig. 1e, f and supplementary table S1, Supplementary Material online), comprises a family of proteins in vertebrates with roles in neutrophil development by mediating repression of TP53 (Keightley et al. 2017). The similarity of the integrase-like domain and the integrase domains of retroviruses and retrotransposons was previously noted, and one of its amino acid residues responsible for the coordination of the metal ion is required for TP53 repression (Keightley et al. 2017). Our results align with the reported evolutionary conservation of this protein family in vertebrates, as 808/828 (97.6%) BTB/POZ-Integrase zinc-binding fusions are found in the phylum Chordata, sharing similar domain architectures (Fig. 1f): an N-terminal viral integrase domain, followed by a BTB/POZ domain, and C-terminal zinc finger arrays. Our approach also identified previously described SETMAR proteins (Robertson and Zumpano 1997; Cordaux et al. 2006; Tellier and Chalmers 2019) with SET and Mariner transposase domains (supplementary table S1, Supplementary Material online). Thus, our approach can identify known instances of co-occurrence of TE- or virus-derived domains with multicopy domains.

Among the remaining top five co-occurring domain fusions, two are phylogenetically restricted to specific taxa: (i) The F-box (IPR001810) and Tc1/mariner Helix-turn-Helix (HTH, IPR041426) fusion is almost exclusively found in nematodes (496/497, 99.8%, see more information below); and (ii) the ankyrin repeat (IPR036770) and DUF4219 (IPR025314, a domain associated with retrotransposons and retroviruses) fusion is restricted to the Streptophyta plant clade (362/362, 100%). The proteins within each of these groups have similar domain architectures (Fig. 1f and supplementary fig. S1d, Supplementary Material online). The taxonomic specificity and the consistent domain structures suggest that the F-box-HTH fusion likely originated from a single fusion event in nematodes, while the ankyrin repeat-DUF4219 fusion likely arose from a single fusion in plants. Conversely, the ankyrin repeat-peptidase A2A (IPR036770-IPR001995) and protein kinase-reverse transcriptase (IPR011009-IPR000477) fusions have a broad taxonomic distribution and divergent domain structures across clades, suggesting that these groups might reflect multiple independent fusion events (Fig. 1e and supplementary fig. S1e, f, Supplementary Material online).

In our analysis of domain structures across groups of proteins, we noticed that in three of the top five co-occurring domain fusions, the TE/virus-derived domain tended to be positioned N-terminally relative to the multicopy domain (Fig. 1f and supplementary fig. S1d, Supplementary Material online). To determine whether this pattern was a general trend across all the 12,803 screen hits, we analyzed domain positions and found an overrepresentation of domains from

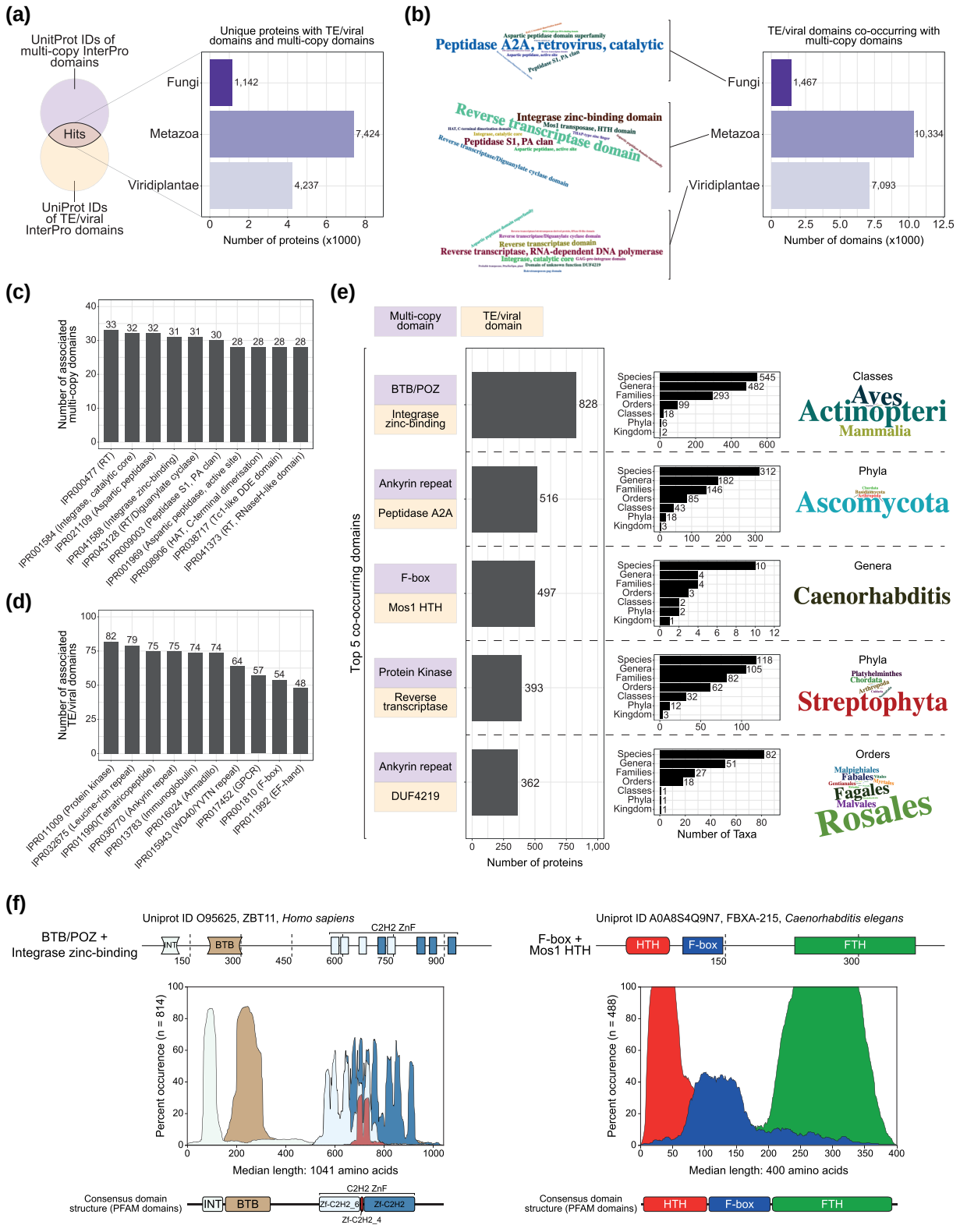

Fig. 1. Identification of eukaryotic proteins with TE/virus-derived and multicopy domains. a) Simplified representation of the approach used to identify co-occurrence of multicopy proteins domains with protein domains of TE/viral origin. Bar plot depicts the number of unique proteins with protein domains of TE/viral origin in major eukaryotic clades. b) Number of TE/viral protein domains co-occurring with multicopy domains in major eukaryotic clades. Word clouds show the top 10 TE/viral domains in each clade. c) The 10 multicopy domains with most associated TE/viral protein domains. d) The 10 TE/viral domains with most associated multicopy domains. e) Largest sets of proteins with the same co-occurring protein domains (bar charts on the left), along with information on their taxonomic distribution (bar charts in the middle and word clouds on the right). Word clouds highlight key taxa where these protein fusions exist. Note that different taxonomic levels are highlighted for each distinct fusion. f) Protein domain architecture of proteins with co-occurring domains in the largest (BTB/POZ-Integrase zinc binding) and the third largest set of proteins (F-box-Mos1 HTH). On top, specific representative proteins are shown, whereas the bottom panels show the consensus protein domain distributions of groups of proteins sharing co-occurring domains, plotted along the median length of these proteins. FTH, FOG-2 homology domain; HTH, Helix-turn-helix; INT, integrase; RT, reverse transcriptase; Znf, zinc fingers.

TEs/viruses in the N-terminal region of these proteins, compared with multicopy domains (supplementary fig. S1g, Supplementary Material online). However, this signal is largely driven by the 1,894 proteins of the top five co-occurring domain fusions with the integrase zinc-binding, Mariner HTH, DUF4219, and reverse-transcriptase domains (supplementary fig. S1h, Supplementary Material online). These findings suggest a bias for TE/virus-derived coding sequences at the N-terminus, driven by four of the top five most frequently co-occurring domain fusions identified in the screen.

An F-box Gene Family With a TE-Derived Helix-Turn-helix Domain in the *Caenorhabditis* Genus

The third largest set of proteins with the same co-occurring domains has F-box (IPR001810) and Tc1/mariner HTH domains (IPR041426) structurally related to the N-terminal DNA-binding HTH domain of the Mos1 transposase of Drosophila mauritiana (Fig. 1e, f) (Richardson et al. 2009). This large set of 497 proteins is phylogenetically restricted to the *Caenorhabditis* genus in nematodes (Fig. 1e), except for four proteins (three in the *Ascaris* and *Heligmosomoides* nematode genera and one in phylum Arthropoda) that do not share an identical domain structure with the remaining 493 proteins (supplementary table S1, Supplementary Material online).

F-box domain-containing proteins adopt a variety of cellular functions, but are mostly known for their roles in the context of Skp, Cullin, F-box (SCF) E3 ubiquitin-ligase complexes (Kipreos and Pagano 2000; Skaar et al. 2013). In specific, F-box proteins interact with protein substrates and bring them in close proximity to SCF complexes allowing substrate poly-ubiquitination and subsequent proteasomal degradation. A detailed study on gene families of ubiquitin-ligase adapters in *Caenorhabditis* has categorized *Caenorhabditis* F-box genes into three families: A1, A2, and B (Thomas 2006). Family A1 genes have F-box and FTH domains, whereas family B genes have F-box and FBA2 domains (Fig. 2a and supplementary table S1, Supplementary Material online). Family A2 was defined by having F-box and FTH domains plus an additional N-terminal domain, which was noted to be related to mariner transposases (Fig. 2a and supplementary table S1, Supplementary Material online). The proteins we identified with co-occurring F-box and Tc1/mariner HTH domains (Fig. 1e, f) correspond to the A2 family of F-box genes in the *Caenorhabditis* genus (Fig. 2a). As the evolutionary history and functional roles of A2 F-box genes were not previously explored (Thomas 2006; Feschotte et al. 2009), we set out to do so.

Given the ubiquity of HTH folds, we confirmed that the structure of the HTH domain of A2 family proteins is TE-derived and not related to other HTH folds. To do this, we performed structural alignments between proteins with TE-derived HTH folds of the same type, i.e. simple tri-helical type (Aravind et al. 2005). An excellent structural alignment is obtained with the HTH domains of A2 family proteins, SETMAR, and Mos1 (Fig. 2b and supplementary fig. S2a–c, Supplementary Material online, all-atom root mean square deviation, RMSD, between 0.7 and 2.1 Å). Conversely, alignment with non-TE-derived HTH folds of the simple tri-helical type is subpar (supplementary fig. S2d–f, Supplementary Material online, RMSD between 2.8 and 6.4 Å). This supports a TE origin for the HTH domain of A2 family proteins.

Proteins with F-box domains are ubiquitous in eukaryotes, including animal genomes (Fig. 2c). However, some nematode clades stand out with an expanded repertoire of F-box genes (Fig. 2c) (Thomas 2006; Röseler et al. 2022). We revisited the composition of the F-box gene families in *Caenorhabditis nematodes*, following their initial categorization (Thomas 2006; Li and Rödelsperger 2022). F-box genes with the domain structure of the A1 and B families of F-box genes are present only in the *Caenorhabditis* genus, suggesting these specific domain architectures are molecular innovations of this nematode genus (Fig. 2c). Conversely, the A2 family is present only in the Elegans group within the *Caenorhabditis* genus, indicating that the HTH domain was captured by the F-box superfamily in the ancestor of the Elegans group, roughly 20 million years ago (Cutter 2008). It is relevant to note the three families only include a fraction of all *Caenorhabditis* F-box domain-containing genes. These observations suggest that a more comprehensive classification of the F-box superfamily using a broader taxonomic sampling might be needed. Phylogenetic analysis of *C. elegans* F-box proteins shows that B family genes form a monophyletic clade, while the A1 family has a patchier distribution (Fig. 2d and supplementary fig. S2g, Supplementary Material online). Although the overall tree topology is not well supported, all the A2 family members fall into a subtree within the larger A1 family, in agreement with a single origin for the A2 family (Fig. 2d). Individual genes within the A2 subtree seem to have lost the HTH domain. Individual losses of other domains are also observed. For example, *fbxa-197* has lost the F-box domain and *Y37H2A.12* has no FTH domain. Altogether, these results support a single origin of the A2 family, followed by an expansion and subsequent losses of the HTH domain.

The majority of *C. elegans* genes with an annotated HTH domain (IPR041426) overlap with genes annotated as *fbxa* (39/44) and have an N-terminal HTH domain preceding F-box and FTH domains (Fig. 2e, of note both A1 and A2 F-box gene families share the *fbxa* nomenclature). Thus, the same overall domain architecture is maintained across the F-box A2 family. Only 5/44 HTH proteins are not annotated as *fbxa* and only 3/44 have isolated HTH domains (Fig. 2e and supplementary fig. S3a, Supplementary Material online). All *C. elegans* HTH domain-containing proteins encode their HTH domains in one single exon (supplementary fig. S3a, Supplementary Material online). We could not detect any annotated TE sequences or other sequence motifs consistently flanking the exons containing the HTH domains, besides splice site motifs (supplementary fig. S3b, c, Supplementary Material online). The sequence of events following the capture is consistent with a previously proposed model (Cosby et al. 2021), involving TE insertion close to host transcripts, followed by exonization of TE-derived protein-coding sequences by alternative splicing.

Next, we interrogated the identity and origin of the captured TE. It is not clear whether the A2 ancestor received the HTH sequence directly from an endemic or horizontally transferred TE, or from an intermediate gene family that was the original recipient of the HTH domain. Further analysis of *Caenorhabditis* genes with HTH domains showed that this domain typically occurs either in combination with a transposase (IPR001888) or within an F-box context (either with an FTH, an F-box domain, or with both). These observations rule out the possibility of a secondary transition. Next, we searched for the most closely related sequences in species outside the Elegans group and in noncoding regions of the *C. elegans* genome. Phylogenetic analysis showed that one of the most closely related sequences corresponds to an intronic region that was

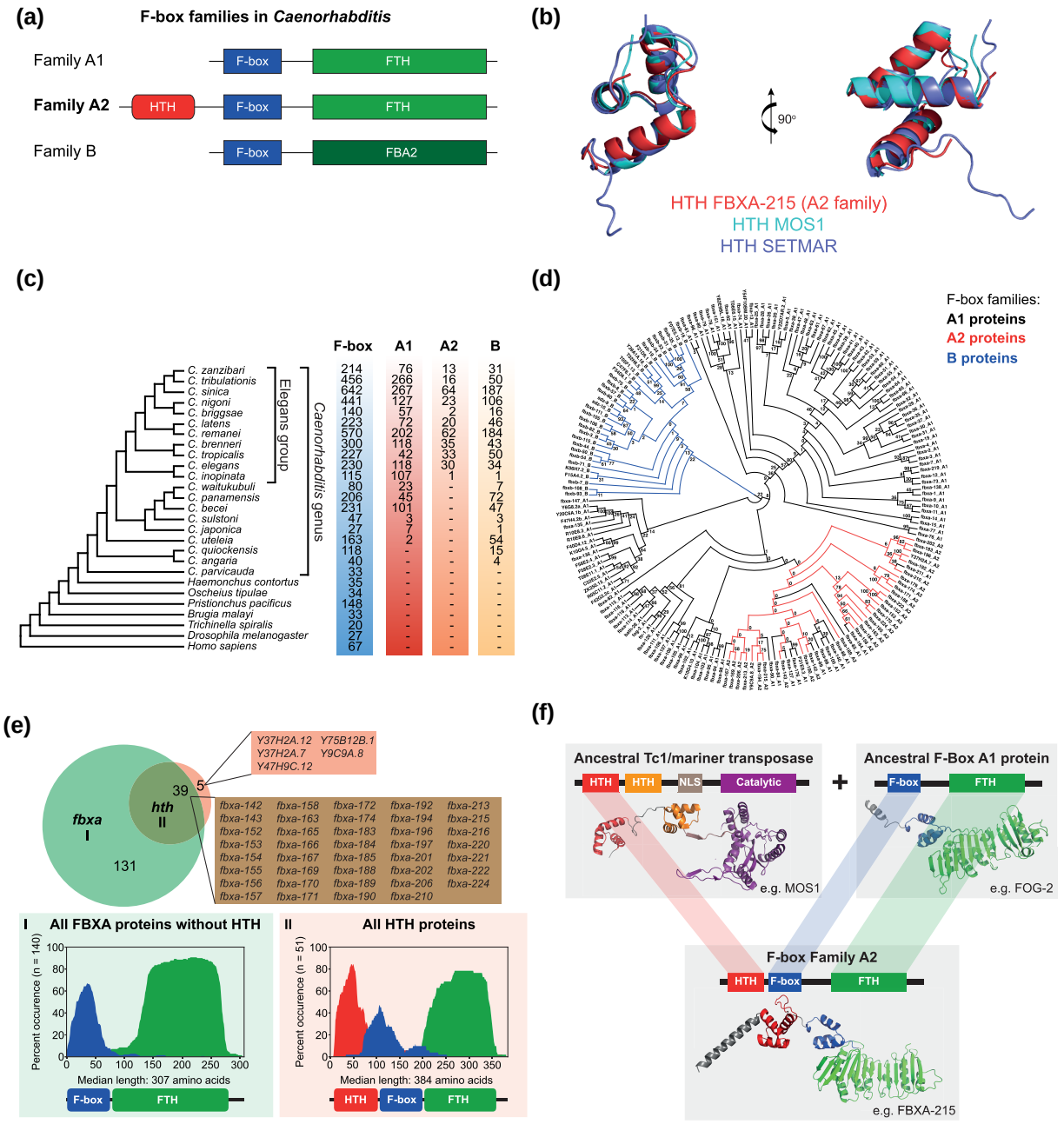

Fig. 2. A subset of the F-box protein superfamily in *Caenorhabditis nematodes* has a Tc1/Mariner-derived Helix-turn-helix domain. a) Schematics illustrating the domain architecture of proteins of the three F-box families in *Caenorhabditis*, as previously categorized (Thomas 2006). b) Structural alignment of known TE-derived HTH domains of *Drosophila mauritiana*'s Mos1 (PDB: 3HOT) and *Homo sapiens*' SETMAR (PDB: 7S03) with a representative HTH domain from FBXA-215, a protein of the F-box A2 family. All these HTH domains are of the simple tri-helical type. c) Genomic data from 27 species (nematodes, *Drosophila melanogaster*, and *Homo sapiens*) were used to date the origin of the A2 family. The cladogram was drawn based on previous phylogenomic analyses (Smythe et al. 2019; Stevens et al. 2019). The species belonging to the *Caenorhabditis* genus and the Elegans group are highlighted in the cladogram. For each species, the total number of F-box domain-containing genes is shown, as well as the members of all three families as in (a). d) An alignment of all F-box proteins from *C. elegans* was used to construct a maximum-likelihood tree. The legend shows the color code for all three gene families, see (a) for a schematic of their protein domain configurations. Numbers at internal branches indicate bootstrap support values (100 pseudoreplicates). e) Venn diagram displaying the overlap between all the genes belonging to the *fbxa* class (including F-box A1 and A2 genes, indicated as I in the figure) according to Wormbase annotations, and all annotated *C. elegans* genes with a Tc1/mariner-derived HTH domain (IPR041426, indicated as II in the figure). The panels below depict the protein structure conservation for all the proteins encoded by the genes in groups I and II (all protein isoforms are included in the analysis). These graphs represent the percent of proteins with specific domains present along amino acid coordinates spanning the median protein length. Consensus protein structure is represented below. f) Model for the emergence of the A2 family of F-box genes in the *Caenorhabditis* genus. In the ancestor of the Elegans group, the N-terminal HTH domain of a Tc1/mariner transposase was captured on the N-terminus of an F-box gene of the A1 family, forming this novel F-box family, which subsequently expanded. Protein structures shown are AlphaFold predictions. FBA2; F-box associated domain 2; FTH, FOG-2 homology domain; HTH, Helix-turn-helix; NLS, nuclear localization signal.

annotated as a Mariner element in the *C. elegans* genome (supplementary fig. S4a–c, Supplementary Material online). This suggests that the HTH domain likely derived from an endemic TE. However, the overall tree topology is poorly supported, and we cannot completely rule out the possibility of an origin from a horizontally transferred TE.

Altogether, these results support a single origin for the A2 family within the *Caenorhabditis* genus, in the common ancestor of the Elegans group, when a Mariner HTH domain was captured by an F-box of the A1 family (Fig. 2f).

F-box A2 Factors Have Signatures of Purifying Selection and a Subset is Expressed in the Germline

Previous estimates of selection on F-box A1 genes identified stronger evidence of positive selection for the FTH domain, which interacts with substrates and positions them near the E3 ligase component of the SCF complex (Thomas 2006). In contrast, the F-box domain showed evidence of purifying selection, consistent with its role in binding to a Skp1 protein, thus connecting the F-box protein to the SCF complex (Thomas 2006). To acquire insights into the function and evolution of the A2 family, we analyzed signatures of selection by examining the coding sequences of their protein domains (Fig. 2a). This analysis supports a predominant signature of purifying selection for the F-box domain, and stronger evidence for amino acid residues under positive selection in the FTH domain (supplementary fig. S5a, Supplementary Material online), in line with previous estimates (Thomas 2006). The HTH domain of F-box A2 proteins displayed a predominant signature of purifying selection (supplementary fig. S5a, Supplementary Material online). In conclusion, the TE-derived HTH domain is predominantly evolving under purifying selection, which, in combination with the fact that it has been maintained in the genome for approximately 20 million years, strongly suggests it may be relevant for the functions of the F-box A2 gene family in the *Caenorhabditis* genus.

To further illuminate the potential functions of F-box factors, including the A2 family, we profiled their expression across development and in adult tissues of *C. elegans* using publicly available RNA-sequencing datasets (Ortiz et al. 2014; Almeida et al. 2019; Serizay et al. 2020). F-box genes are overall lowly expressed throughout development (supplementary fig. S5b, Supplementary Material online), with the F-box B gene family showing higher expression in embryos. In contrast, the A1 family genes are overall more highly expressed than A2 family genes between L1 and L3 larval stages (supplementary fig. S5b, Supplementary Material online). In terms of tissue-specificity in adult animals, >40% of A1 family genes are classified as having intestine-specific expression, whereas ~70% of B family genes are classified as lowly expressed in adults (supplementary fig. S5c, Supplementary Material online). A2 family genes are the most versatile genes, with a similar proportion of genes being assigned as intestine-specific (26.6%), germline-specific (20.0%), and lowly expressed (23.3%, supplementary fig. S5d, Supplementary Material online). The expression of A2 family genes in the germline led us to further explore expression of F-box genes in the germline. 19/116 A1 genes (16.4%), 12/30 A2 genes (40.0%), and 5/34 B genes (14.7%) had detectable expression in oogenic and/or spermatogenic gonads of *C. elegans* (supplementary fig. S5d, Supplementary Material online). 15/19 germline-expressed A1 genes are expressed in both oogenic and spermatogenic gonads, while 4/5 B genes are expressed predominantly in oogenic gonads (supplementary fig. S5e, Supplementary Material online). Again, A2 family genes display the most diverse expression patterns, with a similar proportion of genes predominantly expressed in gonads of each sex and in both sexes (supplementary fig. S5e, Supplementary Material online).

In summary, the predominant signature of purifying selection on the TE-derived HTH domain of F-box A2 genes, along with their expression in the germline, suggest relevant functional roles in the germline.

The HTH Domain of F-box A2 Family Proteins is not Involved in Transcriptional Regulation

A common denominator of the known roles of F-box proteins is the versatile mediation of protein–protein interactions, bridging different proteins or protein complexes (Kipreos and Pagano 2000; Skaar et al. 2013). Given the association of such a versatile protein module with a potentially DNA-binding HTH domain, we reasoned that F-box A2 proteins may have evolved a role in transcriptional regulation of TEs and/or endogenous genes, similar to other host-transposase fusion genes (Bailey et al. 2012; Gray et al. 2012; Tellier and Chalmers 2019; Cosby et al. 2021). As 40% of A2 genes are expressed in the germline (supplementary fig. S5e, Supplementary Material online), such regulation could take place in germline tissues, which comprise a major stage of genetic conflict between TEs and their animal hosts (Ozata et al. 2019; Almeida et al. 2022). Among F-box A2 genes, *fbxa-192*, *fbxa-210*, and *fbxa-215* are the most highly expressed in the germline and in embryos, with *fbxa-215* being the most abundant (supplementary fig. S5b and d, Supplementary Material online). These genes were selected to further interrogate the function of germline-expressed F-box A2 genes. We created mutant strains with complete or partial deletions of the coding sequence of these genes (supplementary fig. S5f, Supplementary Material online). Furthermore, we endogenously tagged FBXA-192 and FBXA-215 with an N-terminal GFP.

Three lines of evidence argue against a transcriptional regulatory role of F-box A2 proteins. First, we profiled mRNA expression in embryos and young adults of wild-type N2 worms and *fbxa-215* mutants and did not find TE families differentially expressed between the two strains, in a consistent manner across developmental stages and growth conditions (supplementary fig. S6a–d and table S2, Supplementary Material online). Similarly, only a small number of protein-coding genes is differentially expressed in *fbxa-215* mutants and these changes are not consistent across developmental stages and growth conditions (supplementary fig. S7a–f and table S2, Supplementary Material online). Second, electrophoretic mobility shift assays failed to detect DNA binding of F-box A2 proteins and their HTH domains to the inverted repeat sequences derived from or similar to their ancestral Mariner TE (supplementary fig. S8a-g, Supplementary Material online). Third, FBXA-192 and FBXA-215 endogenously tagged with GFP display a dispersed localization in the adult germline, with no nuclear localization detected (supplementary fig. S9a, b, Supplementary Material online).

We conclude that F-box A2 genes with high germline expression (FBXA-192/210/215) are unlikely to have a clear, broad transcriptional regulatory role. However, we cannot fully exclude the possibility that F-box A2 genes mediate more subtle, tissue- or cell-type-specific transcriptional regulation of targets unrelated to their ancestral Mariner TE.

F-box A2 Family Factors Have Roles in Fertility and Germline Proteostasis

We further explored the localization of F-box A2 proteins and observed GFP::FBXA-215, but not GFP::FBXA-192, localizing to perinuclear germ granules in embryos (Fig. 3a,

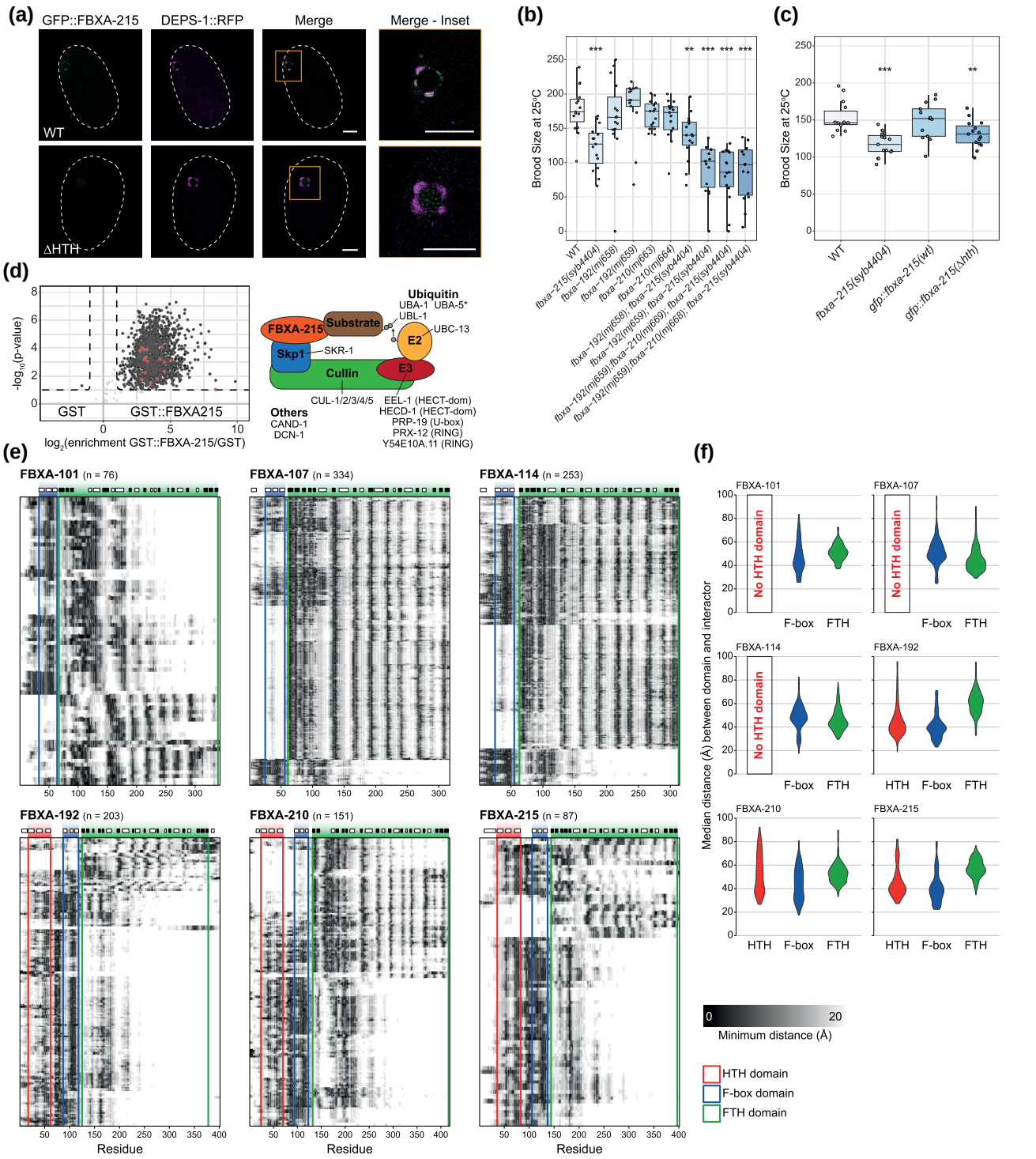

Fig. 3. Germline-enriched A2 F-box proteins in *C. elegans* are required for fertility and function in the context of SCF complexes. a) Confocal photomicrographs illustrating the localization of GFP::FBXA-215 (green), DEPS-1::RFP (magenta), and their co-localization in *C. elegans* embryos. Upper row of panels illustrates the localization of these proteins in animals carrying a wild-type copy of GFP::FBXA-215. Conversely, the lower row of panels shows the localization of these proteins in animals with a mutated GFP::FBXA-215, which lacks the TE-derived HTH domain. Insets focus on the P4 blastomere primordial germ cell (prior to its division into Z2 and Z3 primordial germ cells). Scale bars indicate 10 µm. Images represent single confocal planes images. Dashed lines represent the embryo outline. b-c) Live progeny of the indicated strains at 25 °C. Asterisks and *P*-values assessed by Mann–Whitney and Wilcoxon tests comparing wild-type (WT) N2 worms with the other strains. Horizontal lines represent the median, the bottom and top of the box represent the 25th and 75th percentile. Whiskers include data points that are less than 1.5× interquartile range away from the 25th and 75th percentile. b-c) Significance levels are indicated as follows: *0.01 ≤ *P*-value < 0.05; **0.001 ≤ *P*-value < 0.01; *** *P*-value < 0.001. d) Volcano plot shows proteins pulled-down with GST::FBXA-215 over the GST alone, as identified by label-free quantitative mass spectrometry. Schematics on the right represent the SCF complex factors detected in the interactome of FBXA-215. These factors are highlighted in red in the volcano plot. See detailed results in supplementary table S3, Supplementary Material online. Asterisk indicates proteins that were classified as an interactor of the HTH domain, but not of the full-length protein. e) Heatmaps showing the minimum distance, in Å, between each amino acid residue of the F-box protein bait and the high-confidence interactors. Each row of the heatmap corresponds to one interactor. The domains of the F-box protein bait are indicated as colored boxes, see color key on the bottom right. The secondary structure of each F-box protein is shown on top of the respective heatmap, with the white and black boxes representing alpha helices and beta sheets, respectively. f) Violin plots showing the median distance, in Å, between HTH, F-box, or FTH domains of F-box proteins to the high-confidence protein interactors obtained from the AlphaFold2 multimer screen.

supplementary fig. S9a and c, Supplementary Material online), as evaluated by co-localization with DEPS-1 (average Pearson's co-localization coefficient 0.48 ± 0.21), a factor known to localize to germ granules (Spike et al. 2008; Suen et al. 2020; Huang et al. 2024). Importantly, in-frame deletion of the HTH domain disperses GFP::FBXA-215(ΔHTH) from germ granules (Fig. 3a and supplementary fig. S9d, Supplementary Material online, average Pearson's co-localization coefficient 0.08 ± 0.04), indicating the TE-derived HTH domain is required for the localization of GFP::FBXA-215 to germ granules. In *fbxa-215* mutant animals, germ granules look similar to wild-type, suggesting FBXA-215 is not required for germ granule condensation (supplementary fig. S9e, Supplementary Material online).

Factors localizing to germ granules are required for fertility (Kawasaki et al. 1998; Spike et al. 2008). Given the localization of FBXA-215 to germ granules, we tested whether fertility is impacted by disruption of FBXA-215 and other germline-expressed F-box A2 genes. *fbxa-215* mutant animals display a mild fertility defect at 25 °C when compared to wild-type, whereas *fbxa-192* and *fbxa-210* show no fertility defects (Fig. 3b). However, when these genes were mutated in combination with *fbxa-215*, fertility was more severely impacted, showing that F-box A2 genes contribute to fertility synergistically. *fbxa-215* mutants with an in-frame deletion of the TE-derived HTH domain have a fertility defect similar to the null allele, indicating that this domain is necessary for wild-type levels of fertility (Fig. 3c).

To further understand the roles of F-box A2 factors in the germline, we explored their interactome. We expressed and purified F-box A2 constructs in vitro fused with a glutathione S-transferase (GST) solubility tag (overview of constructs in supplementary fig. S10a, Supplementary Material online), incubated the purified proteins with *C. elegans* extracts and performed GST pulldowns followed by mass spectrometry. In line with the roles of F-box proteins in the context of SCF E3 ubiquitin-ligase complexes, the interactome of FBXA-215 includes SCF complex factors (Fig. 3d and supplementary table S3, Supplementary Material online). Two additional F-box A2 proteins, FBXA-192 and FBXA-210, also interact with SCF complex-associated proteins (supplementary fig. S10b, c and table S3, Supplementary Material online). Furthermore, gene ontology analysis of the interactomes of FBXA-215 and additional F-box A2 factors revealed enrichment for proteins related to the ribosome, mitochondria, and unfolded protein response (supplementary fig. S10d and table S3, Supplementary Material online), hinting at roles in stress responses.

Together, these data show that germline-expressed F-box A2 factors have roles in fertility, and that the TE-derived HTH domain of FBXA-215 is necessary to maintain normal fertility. Interactions with SCF complex factors support roles of F-box A2 proteins in proteostasis.

The HTH Domain May Provide a Protein–Protein Interaction Platform to F-box A2 Proteins

We performed GST pulldowns and mass spectrometry on additional germline-expressed F-box proteins, which do not encode an HTH domain (FBXA-101, FBXA-107, and FBXA-114). F-box proteins with an HTH domain tended to interact with a larger number of proteins compared with other F-box proteins (supplementary fig. S10e, Supplementary Material online). In addition, we expressed the HTH domain and the C-terminal fragment of F-box A2 proteins separately (supplementary fig. S10a, Supplementary Material online) and conducted pulldown-mass spectrometry. This revealed that the HTH domains tended to interact with more proteins than the C-terminal fragments (supplementary fig. S10f, g, Supplementary Material online). These observations may suggest that the HTH domains of F-box A2 proteins mediate binding to diverse protein interactors, possibly expanding the post-translational regulatory capability of these F-box proteins.

To explore how the HTH domain may diversify protein–protein interactions of F-box proteins, we explored the protein–protein interfaces of F-box proteins. To do so, we performed AlphaFold2 multimer high-throughput screens using as baits germline-expressed F-box proteins with (FBXA-192, FBXA-210, and FBXA-215) or without (FBXA-101, FBXA-107, and FBXA-114) HTH domains, and modeled their interactions with the entire germline proteome. To attain a higher level of confidence in thresholding predictions, we first defined high-confidence interactors by analyzing a set of true interactors and noninteractors with a benchmarking dataset. By applying these thresholds, we could define high-confidence interactors of F-box proteins (supplementary fig. S11a, Supplementary Material online). SKR-1, the *C. elegans* ortholog of SKP1, emerged as a top interactor, with an observable interface with the F-box domain of all the F-box proteins tested (supplementary fig. S11b, Supplementary Material online). This observation is consistent with our mass spectrometry results (Fig. 3d and supplementary table S3, Supplementary Material online) and with the ability of F-box domains to interact with SKP1 proteins (Kipreos and Pagano 2000; Skaar et al. 2013). According to AlphaFold2 predictions, the high-confidence interactors of HTH-less F-box proteins make contacts preferentially with the C-terminal FTH domain, in line with the FTH domain being the domain responsible for recognition of protein substrates (Fig. 3e, f and supplementary fig. S11c, Supplementary Material online). In F-box A2 proteins, most interactions occur in the N-terminus of these proteins, in the region overlapping the HTH domain (Fig. 3e, f and supplementary fig. S11c, Supplementary Material online). These results agree with the mass spectrometry results (supplementary fig. S10f, g, Supplementary Material online).

Overall, our data suggest that the TE-derived HTH domain provides a binding platform to recognize potential protein substrates, thus expanding the regulatory potential of F-box A2 proteins.

Helitrons Integrate F-box A2 Genes into Stress-Responsive Regulatory Networks

Since results above pointed towards a role of F-box A2 proteins in the stress response (supplementary fig. S10d and table S3, Supplementary Material online), we sought to understand how F-box genes are regulated and integrated with pathways responsive to stress in *Caenorhabditis*. In *C. elegans*, most F-box genes are located within two clusters on the arms of chromosomes II and V, in TE-rich regions (Fig. 4a–c). Transcription factors can bind TE-derived sequences, regulating neighboring genes, or integrating them into host regulatory networks, including networks responsive to stress (Horváth et al. 2017; Lanciano and Mirouze 2018; Almeida et al. 2022; Fueyo et al. 2022). We therefore developed an approach to address whether the genomic regions harboring F-box domain-containing genes are significantly enriched in TEs (supplementary fig. S12a, Supplementary Material online). We observe variable enrichment of DNA,

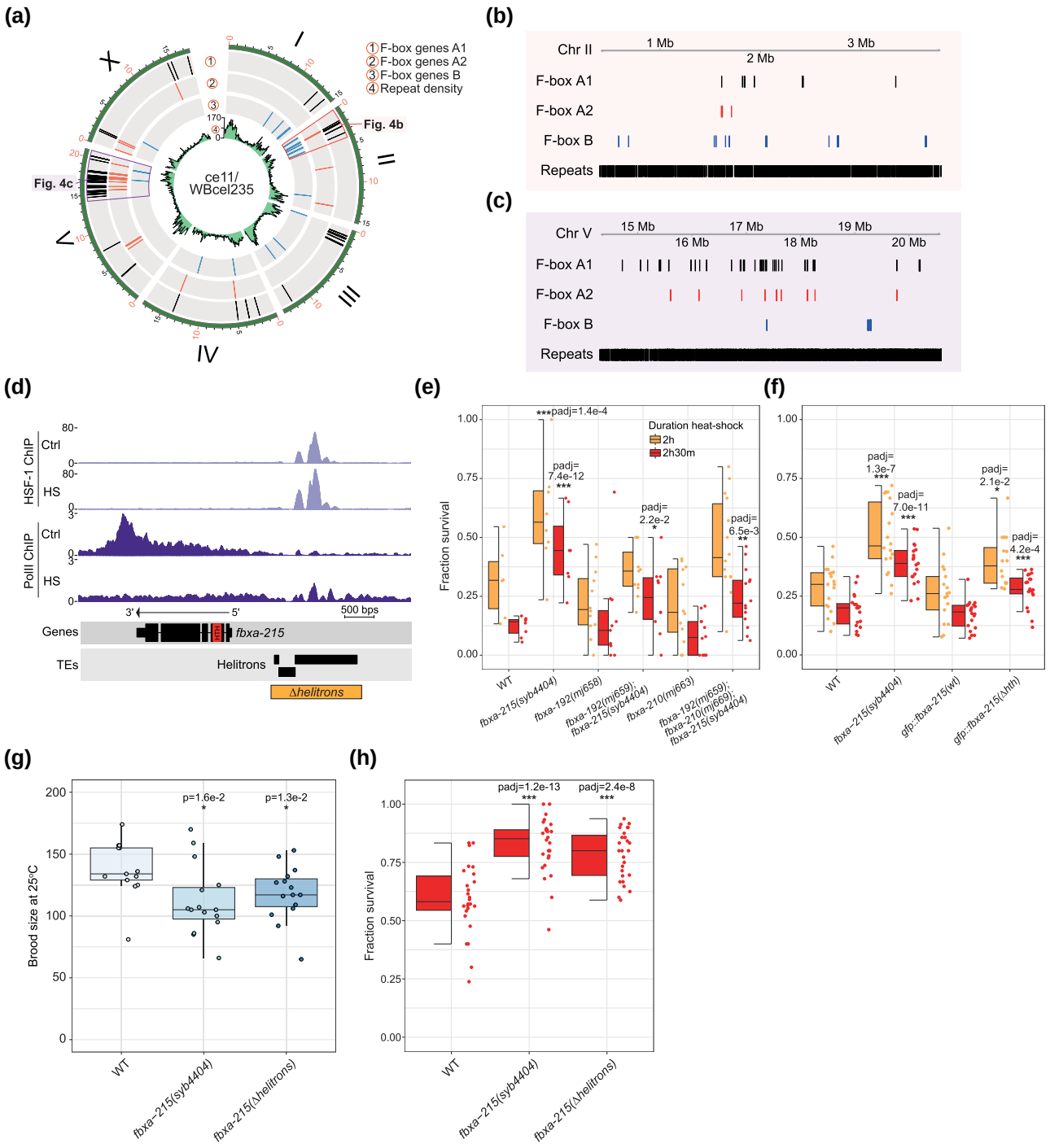

Fig. 4. FBXA-215 modulates thermotolerance downstream of HSF-1-bound helitrons. a) Circos plot showing the location of all classes of F-box genes in the C. elegans genome. Inner track in green displays the repeat density across the genome. b, c) Insets of the major clusters of F-box genes in chromosome II b) and chromosome V c) of *C. elegans*. d) Genome tracks showing HSF-1 and RNA Polymerase II ChIP-sequencing data (Li et al. 2016) in the locus of *fbxa-215*. Ctrl, Control; HS, Heat shock. The HTH domain is indicated in the *fbxa-215* annotation. The range of the Helitron deletion is shown under the TE track as a box. e-f) Worm survival 24 h after 37 °C heat shock for 2 h or 2 h 30 m. Each figure represents two combined experiments. *P*-values show the results of Fisher's exact tests comparing survival of worm strains versus survival of wild-type (WT) N2. g) Live progeny of the indicated strains at 25 °C. Asterisks and *P*-values assessed by Mann–Whitney and Wilcoxon tests comparing wild-type N2 worms with the other strains. h) Worm survival 24 h after 37 °C heat shock for 2 h 30 m. These results represent three combined experiments. Asterisks and *P*-values denote the results of Fisher's exact tests comparing survival of worm strains versus survival of wild-type N2. e–h) Horizontal lines in the box represent the median, whereas the top and top of the box represent the 25th and 75th percentile. Whiskers include data points that are <1.5× interquartile range away from the 25th and 75th percentile.

LTR, and Helitron elements in the vicinity of F-box genes in 11/17 *Caenorhabditis* genomes (supplementary fig. S12b, Supplementary Material online). Specifically, DNA TEs are enriched in 8/17 species, LTRs in 4/17 species, and Helitrons in 3/17 species. This shows that despite general TE enrichment, no single TE class is consistently associated with F-box genes in all species, likely reflecting lineage-specific TE dynamics. We conclude that F-box genes tend to reside in

TE-rich genomic regions, providing opportunity for new regulation to emerge.

In *C. elegans*, we found Helitron TEs directly upstream of *fbxa-215* (Fig. 4d). Helitrons are known to distribute binding sites for Heat shock Factor 1 (HSF-1) in *Caenorhabditis* genomes (Garrigues et al. 2019; Schreiner et al. 2019). HSF1 is a conserved transcription factor in eukaryotes that coordinates transcriptional programmes in a variety of contexts, including development, metabolism, and in response to heat stress (Vihervaara and Sistonen 2014; Gomez-Pastor et al. 2018). In *C. elegans*, HSF-1 has also been shown to regulate developmental programs and heat-stress responses (Chiang et al. 2012; Morton and Lamitina 2013; Brunquell et al. 2016; Li et al. 2016; Edwards et al. 2021). Previously published chromatin immunoprecipation sequencing (ChIP-seq) data (Li et al. 2016) support HSF-1 binding to the Helitrons upstream of *fbxa-215* (Fig. 4d). Moreover, RNA Polymerase II elongation along *fbxa-215* is compromised upon heat shock (Fig. 4d), consistent with previously reported downregulation of *fbxa-215* after heat shock (Brunquell et al. 2016). ChIP-seq and RNA-seq data from soma- and germline-specific depletion of HSF-1 (Edwards et al. 2021) further suggest that HSF-1 binding to the Helitrons near *fbxa-215* is mostly occurring in the germline (supplementary fig. S13a, b, Supplementary Material online).

Next, to investigate the phenotypic effects of the regulation of *fbxa-215* by HSF-1, we quantified survival after heat shock at 37 °C for 2 h and 2 h 30 m. *fbxa-215* mutants displayed enhanced survival after heat shock, when compared with wild-type N2 animals (Fig. 4e). Other F-box A2 mutants did not display this phenotype and did not further enhance the phenotype of *fbxa-215* (Figs. 4e and supplementary fig. S13c, Supplementary Material online). GFP::FBXA-215 did not change subcellular localization upon heat shock (supplementary fig. S13d, e, Supplementary Material online). Of note, in-frame deletion of the HTH domain of FBXA-215 led to increased survival after heat shock compared to wild-type, phenocopying the null mutant and indicating the HTH domain affects thermotolerance (Fig. 4f). Lastly, we determined the phenotypic impact of the Helitrons upstream of *fbxa-215*. To do so, we created mutants with the Helitrons deleted (region deleted indicated in Fig. 4d, not affecting the coding sequence of *fbxa-215*). In these mutant animals, under a nonstressful growth temperature of 20 °C, *fbxa-215* is expressed at wild-type levels (supplementary fig. S13f, Supplementary Material online). Nevertheless, disruption of the Helitrons upstream of *fbxa-215* affects fertility at 25 °C and thermotolerance, phenocopying *fbxa-215* null mutants (Fig. 4g, h).

In summary, we report that F-box genes tend to reside in TE-rich regions across the *Caenorhabditis* genus. Also, Helitrons integrate *fbxa-215* into a thermal stress-responsive regulatory network in *C. elegans*.

An Additional Class of Fusion Proteins With F-box and TE Domains in Zebrafish

The tendency of F-box genes to occur in TE-rich regions provides opportunity for TE capture. This led us to interrogate whether F-box genes may have recurrently captured TE domains throughout eukaryotic evolution. Indeed, in our initial screen, we found 820 F-box domain-containing proteins in eukaryotes with potential fusions with domains associated with TEs or viruses (Fig. 5a, b and supplementary table S1, Supplementary Material online). The majority of these unique proteins (492/820, 60%) correspond to the Mariner HTH domain capture in the *Caenorhabditis* genus that we identified and characterized above (Fig. 5b).

Next, we sought to validate additional instances of F-box domain co-occurrence with TE or viral domains. We identified a set of F-box proteins in fishes with a zinc-chelating domain (CxC7-like cysteine cluster associated with KDZ transposases, IPR041300, Fig. 5c) from KDZ (Kyakuja, Dileera, and Zisupton) TEs (Böhne et al. 2012; Iyer et al. 2014). The latter was hypothesized to be involved in the recognition of specific DNA sequences or chromatin proteins during transposition (Iyer et al. 2014). The CxC7 zinc-chelating domain in question (IPR041300) was defined based on the CxC7 zinc-chelating domain of Zisupton TEs in the platyfish, *Xiphophorus maculatus* (Böhne et al. 2012). The zebrafish (*Danio rerio*) reference genome encodes three genes with F-box and CxC7 domains. The sequence and structure of their proteins confirmed the presence of an F-box domain, and of a TE-derived CxC7 domain (Fig. 5c, d). Lastly, we used RNA-sequencing datasets of zebrafish embryogenesis and early developmental stages (Collins et al. 2012; White et al. 2017) to confirm the domain fusion and interrogate the expression of these genes. We observed reads spanning exons encoding the F-box and CxC7 domains, demonstrating that these two domains are transcribed and spliced together to encode a single polypeptide (supplementary fig. S14a, b, Supplementary Material online). The three F-box::CxC7 genes we identified are expressed in early zebrafish development (Fig. 5e), with at least two of these showing prominent expression in segmentation. These expression patterns hint at developmental roles, which are yet to be explored.

These results, together with the F-box::HTH fusions described above, indicate that F-box domain-containing proteins have captured TE- or virus-derived domains in eukaryotes in at least two occasions.

Discussion

In this work, by exploring a universe of 10 million unique proteins annotated in public databases, we identified over 12,000 proteins in eukaryotes with domains associated with TEs and viruses co-occurring with multicopy protein domains, typically associated with multigene families (Fig. 1; supplementary fig. S1 and table S1, Supplementary Material online). We hypothesized that TE/viral domains are more likely to be retained in gene families with functional redundancy, since the additional copies may buffer the impact of an initially deleterious chimeric transcript. Consistent with this, we observed a positive relationship between the number of proteins with a given domain and the number of instances of co-occurrence with TE/viral domains (supplementary fig. S1b, Supplementary Material online).

We describe the capture of a Mariner HTH domain by a subset of the F-box superfamily (Fig. 2), the F-box A2 family. Our phylogenomic analyses suggest a single capture of the HTH domain by one F-box A1 gene in the ancestor of the Elegans group of *Caenorhabditis*. Given the observed positive correlation between the number of multicopy domains and number of TE/virus fusions (supplementary fig. S1b, Supplementary Material online), the large number of F-box genes in the Elegans group of the *Caenorhabditis* genus (Fig. 2c), and presumably in its common ancestor, may have increased the likelihood of capturing a TE domain. Nevertheless, we have no evidence suggesting recurrent, independent acquisitions of TE-derived HTH domains by

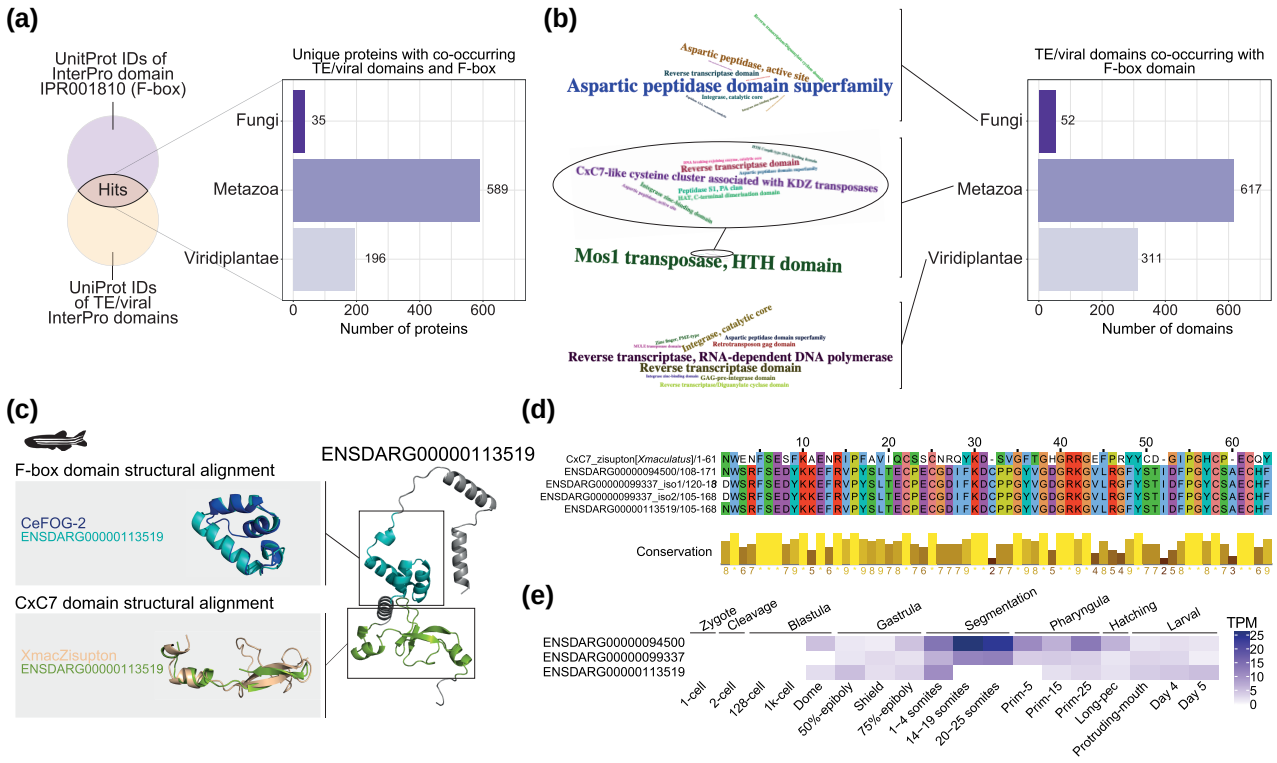

Fig. 5. Additional example of a TE domain capture by F-box proteins in zebrafish. a) Schematic depicting the approach to identify eukaryotic proteins with F-box domains (IPR001810) co-occurring with protein domains of TE/viral origin, a subset of the screen represented in Fig. 1a. Bar plot shows the number of unique F-box proteins with TE/viral domains in major eukaryotic clades. b) Number of TE/viral protein domains co-occurring with the F-box domain (IPR001810) in major eukaryotic clades. Word clouds show the top 10 TE/viral protein domains in each clade. In the metazoan word cloud, given the predominance of the Mos1 transposase HTH domain, mostly corresponding to the HTH capture described in Fig. 2, we show an inset with the remaining terms. c) Example of a protein with an F-box and a TE-derived CxC7-like cysteine cluster domain in zebrafish (*Danio rerio*). This is the protein product of ENSDARG00000113519 one of three zebrafish genes with these protein domains. Upper gray panel shows the structural alignment of the F-box domains of ENSDARG00000113519 and *C. elegans* FOG-2 (CeFOG-2). Lower gray panel shows the structural alignment of the CxC7-like cysteine cluster domain of ENSDARG00000113519 and *Xiphophorus maculatus*' Zisupton. The zebrafish illustration was obtained from phylopic.org. d) Multiple sequence alignment of the CxC7-like cysteine cluster domains of the Zisupton of *Xiphophorus maculatus* (*Xmaculatus*) and the three identified zebrafish genes (both protein isoforms encoded by ENSDARG00000099337 were included). e) Expression of the F-box-KDZ genes identified, in transcripts per million (TPM), during zebrafish early development. Using data from White et al. (2017).

Caenorhabditis F-box genes. Subsequent expansion of the ancestral F-box A2 gene may have been facilitated by the TE-rich nature of the genomic regions of F-box genes (supplementary fig. S12, Supplementary Material online), through recombination or segmental duplications. Supporting this, the two major genomic clusters of F-box genes in *Caenorhabditis* are located in the arms of chromosomes II and V (Fig. 4a–c), regions known for higher recombination rates (Rockman and Kruglyak 2009; Snoek et al. 2019).

TEs had a major role in eukaryotic genome evolution: (i) by providing their protein-coding sequences as building blocks that can be repurposed for other functions; and (ii) by contributing to transcriptional regulation of endogenous genes, sometimes even wiring entire regulatory networks (Bourque et al. 2018; Almeida et al. 2022; Fueyo et al. 2022). The F-box A2 family in *Caenorhabditis nematodes* integrates these concepts, demonstrating both functional and regulatory contributions of TEs. We find striking that F-box A2 genes, which include a TE-derived protein domain with roles in diverse processes, are further integrated into host cis-regulatory networks by unrelated TEs, revealing multiple layers of TE-driven functional and regulatory novelty. For example, *fbxa-215* has Helitron TEs directly upstream of its coding sequence, which are bound by the transcription factor HSF-1 (Fig. 4d). HSF-1 binding upstream of *fbxa-215* has phenotypic implications as deleting *fbxa-215*, its HTH domain, or the upstream Helitrons affects fertility and resistance to heat stress (Fig. 4e–h). These findings underscore the relevance of Helitron-driven genome evolution in the *Caenorhabditis* genus and beyond (Garrigues et al. 2019; Schreiner et al. 2019; Barro-Trastoy and Köhler 2024). These findings also exemplify how stress-responsive TE activity can integrate other genes in stress-responsive processes (Horváth et al. 2017; Lanciano and Mirouze 2018).

The F-box A2 gene family adds to a body of work in nematodes demonstrating that even organisms with a relatively low genomic TE content have a hidden abundance of TE-derived genomic novelties. For example, one recent study reported the discovery of TE-derived structural variation that modulates the expression of factors required for the biogenesis of particular small RNAs, and consequently affects the regulation of targets of these small RNAs (Zhang et al. 2024). Also, a germline cis-regulatory network was unveiled, wired by miniature inverted repeat TEs and a transcription factor related to a transposase (Carelli et al. 2022).

F-box A2 proteins and their HTH domains likely have a variety of biological roles. Here, we focused on three germline-expressed F-box A2 factors, out of a total of 30 A2 genes in *C. elegans*. We found FBXA-192/210/215 have synergistic roles in fertility, but they do not all contribute to

thermotolerance (Figs. 3b, c, 4e, f, and supplementary fig. S13c, Supplementary Material online). This may be due to the fortuitous integration of FBXA-215, but not of FBXA-192/210, in the heat-stress response by Helitrons (Fig. 4d–f). Thus, FBXA-215 seems to be pleiotropic, with potentially unrelated roles in fertility and thermotolerance. The more prominent role of FBXA-215 in fertility may be related to its localization to germ granules, a localization pattern not observed for FBXA-192 (Fig. 3a; supplementary fig. S9a and c, Supplementary Material online). Other F-box proteins with roles in fertility have been described. FBXO24 is expressed in human and mice testes and is required for male fertility in mice (Kaneda et al. 2024; Li et al. 2024). FBXO24 was shown to interact with SCF complex factors and its role in fertility could be via SCF-driven post-translational regulation of target proteins (Kaneda et al. 2024; Li et al. 2024). FBXA-192/210/215 interact with SCF complex factors (Fig. 3d and supplementary fig. S10b-c, Supplementary Material online), raising the possibility that post-translational regulation of specific protein targets underlies the fertility defect.

The proteome-wide prediction and modeling of F-box interactors suggests that the TE-derived HTH domain of F-box A2 proteins may alter protein–protein interactions, and diversify the repertoire of protein substrates that SCF complexes can polyubiquitinate and help degrade (Fig. 3e, f; supplementary fig. S11c, Supplementary Material online). The mechanistic basis should be further explored in subsequent studies. We propose that the co-option of TE-derived protein sequences as protein–protein interaction platforms contributing to post-translational regulation may represent a novel regulatory theme. Accordingly, we identified other domains known to mediate protein–protein interactions, such as BTB/POZ (Bonchuk et al. 2023), co-occurring with TE/viral domains (Fig. 1e, f and supplementary table S1, Supplementary Material online).

We observed a similar proportion of A2 genes expressed in the germline and in the intestine (supplementary fig. S5c, Supplementary Material online). FBXA-158 is an F-box A2 gene expressed in the intestine, which promotes thermotolerance in the context of the intracellular pathogen response targeting Microsporidia and viruses (Panek et al. 2020). The direct genomic vicinity of *fbxa-158* lacks Helitrons (supplementary fig. S3c, Supplementary Material online), suggesting its role in thermotolerance emerged independently of the Helitron-HSF-1 axis in the germline. Other recent studies identified nematode F-box genes at the center of hybrid incompatibility and toxin-antidote systems (Tikanova et al. 2024; Xie et al. 2024). Taken together, F-box genes are emerging as an evolutionarily labile toolkit that nematode genomes maintain and co-opt in diverse circumstances, with an impact in immunity, fertility, and viability. We propose that the variety of TEs enriched in the vicinity of F-box genes may enable their integration in various processes and in distinct cell types, in ways comparable to the integration of FBXA-215 in thermotolerance (Fig. 4a–d and supplementary fig. S12, Supplementary Material online).

In addition to the F-box A2 family in *Caenorhabditis*, we report another class of proteins where an F-box domain co-occurs with a CxC7 zinc-chelating domain in fishes (Fig. 5). This domain is defined by the CxC7 domain of platyfish Zisupton TEs (Böhne et al. 2012; Paysan-Lafosse et al. 2023). Of note, one of the protein isoforms of platyfish Zisupton TEs has been noted to include an F-box domain (Böhne et al. 2012), establishing precedence for the co-occurrence of F-box domains and Zisupton TE domains. The work described here, together with the previously described Zisupton-F-box domain association (Böhne et al. 2012) indicates recurrent association of F-box domains and TE domains. This may reflect recurrent captures of TE domains by F-box proteins, or, alternatively, capture of F-box domains by TEs. While we have not found any evidence supporting the latter scenario, we observed a tendency for F-box genes to locate in TE-rich regions of *Caenorhabditis* genomes (Fig. 4a–c and supplementary fig. S12, Supplementary Material online). It will be useful to determine if this tendency extends beyond nematodes. If so, the residence of F-box genes close to an abundance of TEs may provide ample opportunities for captures throughout evolutionary time. The co-occurrence with TEs may be a general feature of gene families, not just of F-box genes, as gene families were recently shown to locate in TE-rich regions of eukaryotic genomes (Gozashti et al. 2024).

This study underscores the transformative role of TEs in eukaryotic evolution, contributing functional and regulatory novelties that influence complex traits, such as stress responses, immunity, and fertility. Building on our findings and those of other screens (Cosby et al. 2021; Coronado-Zamora and González 2023; Oggenfuss et al. 2024), future research should aim to uncover the full extent to which TE- and virus-derived domains diversify eukaryotic proteomes and generate functional novelty.

Materials and Methods

Identification of Proteins With TE/Viral and Eukaryotic Multicopy Protein Domains

We first compiled and curated lists of InterPro protein domains (Paysan-Lafosse et al. 2023) associated with multicopy genes and domains associated with TEs/viruses (supplementary table S1, Supplementary Material online). Next, we downloaded lists of Uniprot IDs and other metadata associated with these InterPro domains. We cross-referenced both sets of protein domains for shared UniProt IDs and concatenated all the IDs identified, along with relevant metadata (supplementary table S1, Supplementary Material online). The resulting data were imported into R (R Core Team 2021) to generate plots and conduct statistical tests. R scripts available at https://github.com/migueldvalmeida/F-box_TEs. We used the following R packages: tidyverse (Wickham et al. 2019), reshape2 (Wickham 2020), ggrepel (Slowikowski et al. 2023), wordcloud2 (Lang and Chien 2018), and taxonomizr (Sherrill-Mix 2023). Protein structure schematics (upper part of Fig. 1f) were created using DAVI (Saighi et al. 2021), with best match cascade. For BTB/POZ-integrase genes, we used the FASTA sequences of 160 mammalian ZBT11 proteins as DAVI input, while for F-box-Mos1 HTH we used 42 *C. elegans* proteins as input, and arbitrarily chose one representative protein structure schematic for each group of proteins.

Protein Domain Annotation and Phylogenetic Analysis

Comparative genomic data of several nematode species were downloaded from WormBase Parasite (version: WBPS16) (Howe et al. 2017). For *Pristionchus pacificus*, we used the latest gene annotation (version: El Paco gene annotation V3)

(Athanasouli et al. 2020). Protein sequence for *Homo sapiens* and *Drosophila melanogaster* were downloaded from the Ensembl (release-93) and Ensembl Metazoa (release-40) ftp servers. In case of multiple isoforms per gene, the longest protein was chosen as the representative isoform. Protein domain annotation was done using the hmmsearch program of the HMMer package (version 3.3, hmmer.org, option -e 0.001) by searching against the Pfam database Pfam-A.hmm (version 3.1b2, February 2015) (Mistry et al. 2021). F-box containing genes were defined based on the presence of one of the domains PF00646 (F-box), PF12937 (F-box-like), PF13013 (F-box-like_2), PF15966 (F-box_4), PF18511 (F-box_5). The FTH and FBA2 domain were defined by the Pfam profiles PF01827 and PF07735, respectively. For phylogenetic analysis, F-box genes of all three families in *C. elegans* were aligned with the MUSCLE software (version 3.8.31) and the program raxml was run to compute a maximum-likelihood tree (version: 8.2.11, options: -m PROTGAMMAILG -f a) using 100 pseudoreplicates to compute bootstrap support values (Edgar 2004; Stamatakis 2014). To screen for homologous HTH domains, we performed a BLASTP search for *C. elegans* HTH domains against the NCBI nr database excluding any *Caenorhabditis* species and a BLASTP search against *Caenorhabditis* species outside of the Elegans group on WormBase ParaSite. Homologous noncoding sequences were identified by TBLASTN searches against the *C. elegans* genome (version: 2.10.1+, options: -max_target_seqs 2 -evalue 0.001). Phylogenetic analysis was performed as described above.

Transcript Structure Annotation and Sequence Motif Search

The sequences of all HTH domain-containing unspliced transcripts in *C. elegans* were obtained from Wormbase and plotted with Exon–Intron Graphic Maker (http://www.wormweb.org/exonintron), using InterPro domain annotations. Transcripts were categorized according to their exon number and location of the annotated domains. To search for sequence motifs flanking the HTH domains, we manually extracted the noncoding sequences upstream and downstream of the HTH-encoding exon from F-box A2 unspliced transcripts. Then, these sequences were used as input for MEME, a MEME Suite tool (Bailey et al. 2015). Analysis was conducted in classic mode, requesting eight motifs as output.

Selection Analysis

With the approach described above (in the protein domain annotation and phylogenetic analysis methods section), we identified all the F-box A2 genes in the *Caenorhabditis* genomes and extracted their protein-coding and protein sequences from the respective transcriptomes and proteomes (downloaded from WormBase ParaSite, version WBPS16). The protein domain coordinates determined by HMMer were used to isolate the protein-coding and protein sequences of the domains from the full-length sequences. All the protein sequences of the HTH, F-box, and FTH domains were subsequently aligned using MAFFT v7.475 (Katoh and Standley 2013), using option --auto (selected FFT-NS-i strategy). Pal2nal v14 was used to create a reverse alignment from the protein sequence alignments and the respective protein-coding sequences (Suyama et al. 2006). The latter was used as input for selection tests in Datamonkey (Weaver et al. 2018), with Mixed Effects Model of Evolution (MEME) (Murrell et al. 2012). The results were plotted on an R framework (R Core Team 2021), using the following R packages: tidyverse (Wickham et al. 2019), reshape2 (Wickham 2020), patchwork (Pedersen 2023). The MEME method is sensitive to episodic selection and identified strong support for positive selection in 7/53 and 4/43 sites for the HTH and F-box domains (with P-value < 0.05), and 48/244 sites for the FTH domain.

Visualization of Protein Structure

Structural analysis and visualization were conducted in Open-Source PyMOL v2.5 (The PyMOL Molecular Graphics System, v2.5, Schrödinger, LLC). All the structures were downloaded from AlphaFold Protein Structure Database and PDB. Structures from AlphaFold Protein Structure database: AF-O62474-F1 (*Caenorhabditis elegans* FBXA-215), AF-Q9XUX2-F1 (*C. elegans* FBXA-192), AF-E0R7K6-F1 (longest isoform of *C. elegans* FBXA-210), AF-G5EDX9 (*C. elegans* FBXA-101), AF-Q9U2P1 (*C. elegans* FBXA-107), AF-Q9U2X6 (*C. elegans* FBXA-114), AF-G5EBU7-F1 (*C. elegans* FOG-2), AF-G5EGQ5-F1 (*C. elegans* MBR-1), AF-P13528-F1 (*C. elegans* UNC-86), and AF-Q9GR61-F1 (*C. elegans* RBP-10). Structures from PDB: 3HOT (*Drosophila mauritiana* Mos1, obtained by X-ray diffraction) (Richardson et al. 2009), 7S03 (*Homo sapiens* SETMAR, obtained by X-ray diffraction) (Chen et al. 2022), 2M3A (*C. elegans* KNL-2, structure in solution obtained with NMR), 5LUX (H. sapiens CDX1, obtained by X-ray diffraction) (Yin et al. 2017), and 2MGQ (*C. elegans* CEH-37, structure in solution obtained with NMR). For the structural alignments, only the helix-turn-helix domains were used, comprising the following amino acid ranges: FBXA-192 aa16-87, FBXA-210 aa24-76, FBXA-215 aa27-81, Mos1 aa5-55 (aligned only the first HTH), SETMAR aa330-394, MBR-1 aa147-193, UNC-86 aa363-437, RBP-10 aa1-67, KNL-2 aa1-67, CDX1 aa153-215, and CEH-37 aa52-117.

The HTH, F-box, and FTH domains were annotated in the protein structures with the guidance of the domain coordinates from HMM as defined above, and according to extant annotations on Wormbase and Uniprot. Structural alignments of protein domains were done with Open-Source PyMOL v2.5 (The PyMOL Molecular Graphics System, v2.5, Schrödinger, LLC), using the align command. To compare the structural alignment of HTH domains, we calculated all-atom RMSD (Å) values using PyMOL with the super command, cycles = 0. Results were then plotted on R, using packages tidyverse (Wickham et al. 2019) and reshape2 (Wickham 2020). ChimeraX (Goddard et al. 2018) was used to visualize the interaction between SKR-1 and F-box proteins.

The visualization of the consensus protein structure of a group of proteins was produced using DomainViz (Schläpfer et al. 2021) via its web server (https://uhrigprotools.biology.ualberta.ca/domainviz). We used default settings and the FASTA sequences of the groups of proteins indicated as input.

As there were no structures available for fish F-box::CxC7 and for Xiphophorus maculatus Zisupton protein (UniProt ID: G3KL21, used only the CxC7 domain, residues 1333 to 1393 according to InterPro annotation), we used AlphaFold Colab (Jumper et al. 2021) to model their structures (monomer model, run_relax). Amino acid sequences and predictions available at https://github.com/migueldvalmeida/F-box_TEs.

Caenorhabditis elegans Genetics and Culture

Animals were cultured on HB101 bacteria according to standard lab conditions (Brenner 1974). *C. elegans* were grown at 20 °C unless otherwise specified. We used the Bristol strain N2 as the reference wild-type strain. supplementary table S4, Supplementary Material online lists all strains used in this study. Strain design and other analysis were aided by Wormbase resources (Davis et al. 2022).

CRISPR/Cas9 Genome Editing

Mutants were generated using CRISPR/Cas9 genome editing. Cas9 ribonucleoprotein complexes were pre-assembled in vitro and injected into the germlines of wild-type N2 animals, or *fbxa-192(mj659); fbxa-215(syb4404)* double mutants (strain SX3717, see supplementary table S4, Supplementary Material online).

Desalted and deprotected 2′-O-methylated Edit-R guide RNAs (gRNAs) were used (Horizon Discovery Biosciences). These gRNAs are able to bind an Edit-R synthetic tracrRNA (Horizon Discovery Biosciences) in vitro, due to complementary sequences on their 3′ end (Jinek et al. 2012). See supplementary table S5, Supplementary Material online for a list of gRNAs used. The injection mixture comprised 25 mM KCl, 7.5 mM HEPES-KOH pH = 8.0, 1 µg/µl tracrRNA, 160 to 200 ng/µl of each gRNA, and 80 to 160 ng/µl dpy-10 gRNA (co-CRISPR, see below). In some experiments, 5 ng/µl pCFJ104 body wall muscle mCherry reporter was also added to the mix. These components were mixed by gentle tapping, incubated at 95 °C for 5 min and allowed to cool down at room temperature (for at least 5 min). Then, recombinant Cas9 (CP02, PNA Bio) was added to a final concentration of 250 ng/µl, mixed and incubated for 5 min at room temperature. Finally, 20 ng/µl of a DNA dpy-10 repair template was added along with nuclease-free water for a total volume of 10 µl. The mixture was then spun at 21,000 × *g* between 2 and 5 min on a table-top centrifuge. Worms were immobilized on 2% agarose microinjection pads, in Halocarbon oil 700 (Sigma-Aldrich). Injection was performed with InjectMan 4 (Eppendorf) and Eppendorf femtotips using an Olympus IX71 microscope. F1 worms were selected based on the dumpy co-CRISPR phenotype (Kim et al. 2014), due to dpy-10 mutation, and were screened by PCR and Sanger sequencing to identify and confirm deletion alleles. Additional strains were created by SunyBiotech, using CRISPR/Cas9 genome editing (see supplementary table S4, Supplementary Material online). CRISPR–Cas9 mutants used in the experiments shown in this work were outcrossed 4× with wild-type N2 animals.

Brood Size Determination

Done as previously described (Almeida et al. 2018; Dietz et al. 2021). In short, 15 L3 worms of each strain (grown at 20 °C) were individually isolated and grown at 25 °C. After onset of egg laying, animals were transferred to a new plate every day, until no eggs were laid for two consecutive days. Live progeny was counted ∼24 h after removing the parent. The progeny of worms that died before egg laying terminated, for example, by dehydration on the side of plate, was excluded from the analysis. Each experiment was performed at least two independent times. Statistical significance was tested using two-sided Wilcoxon–Mann–Whitney tests.

Thermotolerance Experiments

We performed the following protocol based on previous studies (Panek et al. 2020) and published experimental recommendations (Zevian and Yanowitz 2014). Two gravid adults of each strain were isolated into a fresh plate, allowed to lay eggs for approximately 1 h, then removed off the plate. Eggs laid during this 1-h period were allowed to develop at 20 °C for 48 h producing a synchronized population of L4 stage animals. After 48 h, plates were sealed with parafilm and the L4 worms were heat shocked for 120 or 150 min at 37 °C. After heat shock plates were moved to room temperature for 30 min, and then to a 20 °C incubator. Worms were scored for survival ∼24 h after heat shock ended. Worms were scored as dead when they failed to respond to touch and did not show any pharyngeal pumping. Each experiment was performed at least two independent times. Statistical significance was tested using Fisher's exact test.

In Vitro GST Fusion Protein Expression and GST Pulldowns

Expression of GST Fusions in *E. coli*

Constructs were cloned, expressed and purified as previously reported (Almeida et al. 2018). Plasmids encoding GST-F-box fusion proteins were retransformed in Rosetta 2(DE3) Singles Competent Cells (Novagen, 71400-3), as per the manufacturer's instructions. One colony was inoculated in a 5 ml preculture, supplemented with Ampicillin and Chloramphenicol (100 µg/ml and 25 µg/ml, respectively), and grown overnight at 37 °C. This preculture was used to inoculate a 200 to 250 ml culture of LB, supplemented with Ampicillin and Chloramphenicol (100 µg/ml and 25 µg/ml, respectively) and grown at 37 °C, up to an OD_{600} of 0.5 to 0.9. Then, protein expression was induced with 1 mM isopropyl β-D-thiogalactoside (IPTG, Melford, I56000) and incubated overnight at 18 °C. Next, bacteria were collected by centrifugation (4 °C, 4000 rpm, 15 to 30 min) and pellets were frozen at −80 °C.

GST-on Bead Purification

Bacteria pellets were resuspended in Lysis Buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1 mM DTT, 0.1% Triton-X, and protease inhibitors, cOmplete Mini, EDTA-free, Roche, #4693159001). Next, the samples were sonicated with a Sonics Vibra-Cell instrument four times for 2 min (amplitude = 30%), with 5–10 min pauses between cycles. Debris were pelleted by centrifugation (4 °C, 4,000 rpm, 30 to 45 min) and discarded, supernatant was filtered through a 0.20 µM Sartorius Minisart filter. A 250 to 300 µl slurry of Glutathione Sepharose High Performance beads (GE Healthcare, 17527901) was washed three times with 1 ml lysis buffer. Centrifugation steps in washes were conducted at 800 × *g* for 3 min. The cell lysate was then incubated with the beads between 2 and 3 h at 4 °C with end-to-end mixing. After incubation, beads were washed between 10 and 15 times with wash buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1 mM DTT, and protease inhibitors, cOmplete Mini, EDTA-free, Roche, #4693159001) and, after the last wash, suspended 1:1 in wash buffer. 5% glycerol was added to the proteins for short-term storage at 4 °C. To examine the purity of the preparation, 1× LDS buffer (Life Technologies, #NP0007) was added to the samples, and samples were subsequently loaded onto a denaturing NuPAGE Bis-Tris 4% to 12% gel

(Life Technologies, #NP0335BOX) in a NuPAGE MOPS SDS running buffer (Life Technologies, #NP0001) and ran at a constant voltage of 150 V. Gels were stained with InstantBlue Coomassie stain (Abcam, #ab119211) and imaged.

Worm Extract Preparations

A synchronized population of young adult worms was obtained by bleaching gravid adults grown at 20 °C to obtain embryos, allowing the embryos to hatch overnight in M9, with gentle mixing, bringing the synchronized L1s on plate, and growing them at 20 °C. These synchronized animals were collected and lysed in Lysis Buffer (25 mM Tris HCl pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 1 mM DTT, 0.1% Triton X-100 and protease inhibitors: cOmplete Mini, EDTA-free, Roche, #4693159001). Lysis was performed by sonication (10 to 20 cycles of 30 s ON and 30 s OFF) in a Bioruptor Plus (Diagenode). For embryo collection, large populations of gravid adults grown at 20 °C were bleached, embryos were thoroughly washed with M9 buffer, and frozen in Lysis Buffer using liquid N2. Lysis was performed by grinding frozen embryo pellets and douncing with 40 strokes, piston B. Lastly, lysates were cleared by centrifugation (15 min at 21,000 × *g*, 4 °C) and protein concentration was measured using Bradford Protein assay according to manufacturer's instructions (Bio-Rad, #5000006).

GST Pulldowns

Five micrograms of GST-F-box beads were washed three times with pulldown wash buffer (25 mM Tris HCl pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 1 mM DTT and protease inhibitors: cOmplete Mini, EDTA-free, Roche, #4693159001). Subsequently, 300 to 700 µg of worm extract (either from young adult animals or embryos) was incubated with the beads for 3 h, at 4 °C, with end-to-end mixing. Beads were washed three times (with pulldown wash buffer) and resuspended in a total 25 µl volume including 1× LDS buffer (Life Technologies, #NP0007) and 100 mM DTT. Lastly, beads were boiled at 95 °C for 10 min and frozen at −20 °C until processing of samples for mass spectrometry took place (see below).

Mass Spectrometry

In-Gel Digest

In-gel digestion followed previously established procedures (Shevchenko et al. 2007). Samples underwent electrophoresis on a 4% to 12% Bis-Tris gel (NuPAGE, Thermo Scientific, #NP0321) for 8 min at 180 V in 1× MES buffer (Thermo Scientific, NP0002). Each lane was excised and cut into ∼1 mm × 1 mm pieces using a sterile scalpel and transferred into one well of a 96 well hydrophilic low protein binding filter plate (Merck Millipore, MSBVN1210). Gel pieces were destained using destaining buffer (50% 50 mM ammoniumbicarbonate buffer, or ABC, pH 8.0, 50% ethanol) at 37 °C with vigorous agitation until completely destained. Subsequently, gel pieces were dehydrated by immersing them in 100% acetonitrile for 10 min at 25 °C with shaking. The gel pieces were incubated in reduction buffer (50 mM ABC, 10 mM DTT) at 56 °C for 60 min. The reduction was followed by incubation in alkylation buffer (50 mM ABC, 50 mM iodoacetamide) for 45 min at room temperature in the dark. Gel pieces were washed with digestion buffer (50 mM ABC) for 20 min at 25 °C. Then, gel pieces were dehydrated again by incubation in pure acetonitrile until gel pieces got white and hard. Samples were further dried at 80 °C until the filter membrane of the 96 well plate turned white. The dried gel pieces were rehydrated in a trypsin solution (50 mM ABC, 1 µg MS grade trypsin per sample, Serva Electrophoresis 37286) and the filter plate placed on top of a 96 deep well collection plate (Eppendorf, 951032603). Gel pieces were incubated overnight at 37 °C. The plate assembly was centrifuged at 300 × *g* for 2 min. Flowthrough containing tryptic peptides was collected and combined with additional elutes obtained by treating the gel pieces with extraction buffer (50 mM ABC, 30% acetonitrile) twice and a further step with pure acetonitrile for 10 min at 25 °C, shaking at 300 rpm. The dehydration step with acetonitrile was repeated until gel pieces got white and hard. The sample containing tryptic peptides was reduced to 10% of the original volume using a Concentrator Plus (Eppendorf, #5305000304, settings V-AQ) to remove acetonitrile and purified using the StageTip protocol.

Stage Tip Purification

Stage tip purification followed mainly a previously described method (Rappsilber et al. 2007). Desalting tips were created by stacking two layers of C18 material (Affinisep AttractSPE, #SPE-Disk-Bio-C18-100-47.T1.20) within a 200 µl pipet tip. These tips were primed using pure methanol. Following activation, they underwent successive rinses with solution B (80% acetonitrile, 0.1% formic acid) and solution A (0.1% formic acid) each for 5 min before application of the tryptic peptide samples. Subsequently, a final wash with solution A was performed. When utilized, peptides were eluted using solution B. The resultant samples were centrifuged in a Concentrator Plus for 10 min to remove acetonitrile and were adjusted to a volume of 14 µl using solution A.

MS Analysis

For MS analysis a volume of 5 µl desalted and eluted peptides of each sample were injected and separated on a nanocapilary column (New Objective, 25 cm long, 75 µm inner diameter) packed in-house with C18 (Dr. Maisch GmbH) for reverse-phase chromatography. This setup was interfaced to an EASY-nLC 1000 system (Thermo Scientific) coupled to a Q Exactive Plus mass spectrometer (Thermo Scientific). Peptides were eluted from the column employing an optimized 2-h gradient ranging from 2% to 40% of 80% MS grade acetonitrile/0.1% formic acid solution at a flow rate of 225 nl min^{1}. The mass spectrometer was operated in a data-dependent acquisition mode, conducting one MS full scan followed by up to 10 MS/MS scans using HCD fragmentation. All raw files were processed using MaxQuant (version 2.4.2.0) (Cox and Mann 2008) and matched against the *C. elegans* Wormbase protein database (version WS269, 60,000 gene transcripts) and the Ensembl Bacteria *E. coli* REL606 database (version from September 2018, ASM1798v1, 4,533 gene transcripts) for proteins originating from the *E. coli* feeding strain. Carbamidomethylation (Cys) was set as a fixed modification, while methionine oxidation and protein N-acetylation were considered as variable modifications. Enzyme specificity was set to trypsin with a maximum of two miscleavages. LFQ quantification without fast LFQ was performed, requiring at least 2 LFQ ratio counts, and the match between run option was activated. Filtering and analysis were done in R (R Core Team 2021). Interactors of GST::F-box fusion proteins were defined by significant

enrichment over GST control. For GST pulldowns from young adult extracts, interactors were defined by fold change > 2 and *P*-value < 0.1. For GST pulldowns from embryo extracts, interactors were defined by fold change > 1.5 and *P*-value < 0.1. Gene ontology was conducted on sets of F-box protein interactors using g:Profiler web server's g:GOST functional profiling with the *Caenorhabditis elegans* genome and default parameters, including driver terms (Kolberg et al. 2023). See detailed mass spectrometry and gene ontology results in supplementary table S3, Supplementary Material online.

Electrophoretic Mobility Shift Assay

After purification (see above), GST fusion proteins to use in EMSA were eluted from the beads five times with 150 µl of elution buffer (wash buffer with 10 mM reduced glutathione, pH readjusted to 7.5 with NaOH). The eluates were combined, and the sample concentrated and buffer-exchanged back to wash buffer on a 3k Amicon Ultra-0.5 ml centrifugal filter (Merck Millipore). Protein concentration was quantified by measuring OD_{595} in a Bradford Protein Assay (BioRad, #5000006) relative to bovine serum albumin standards. Five micrograms of the sample were run on a denaturing gel to examine the purity of the preparation and correct protein size.

PAGE-purified single-stranded DNA oligonucleotides (supplementary table S5, Supplementary Material online) were annealed by incubation in a heating block set at 89 °C and allowed to cool down to room temperature overnight. The annealing reaction comprised 5 µl of each 100 µM complementary oligonucleotide, 2 µl annealing buffer (200 mM Tris–HCl pH 7.5, 100 mM MgCl2, 1 M KCl) and 8 µl nuclease-free water. Binding reactions were carried out in wash buffer (50 mM Tris–HCl pH 7.5, 150 mM NaCl, 1 mM DTT, EDTA-free protease inhibitors Roche, 2.5% glycerol) in a total volume of 15 µl with purified protein and 77.3 nM DNA. After incubating at 25 °C for 15 min, 2.5 µl of loading dye (50 mM Tris–HCl pH 7.5, 150 mM NaCl, 33% glycerol, 0.1% bromophenol blue) was added. The samples were run on a native 1.5% UltraPure agarose (Invitrogen) gel in 1× TBE pH = 7.5 for 80 min (4 °C, 90 V constant voltage). The gel was stained with 1× SYBR Gold (Invitrogen) in 1× TBE pH = 7.5 for 40 min rocking, and then imaged.

Microscopy

A flat 2% agarose pad was prepared on a glass slide. To prepare gravid adult animals for imaging, animals were picked into a droplet of M9 to wash off the excess *E. coli* bacteria the worms grow on. Subsequently, animals were transferred to another droplet of M9 on the agarose pad, supplemented with 2.5% to 5% sodium azide for immobilization. Embryos in utero, or extruded during the immobilization process were also imaged. This approach was employed to create the images in supplementary fig. S9, Supplementary Material online, using a Nikon Ti2-E inverted microscope (NIS-Elements AR image acquisition software, v5.42.06) equipped with a Plan Apo VC 20x/0.75 air objective, a Plan Apo VC 60x/1.2 water objective, a Nikon DS-Qi2 camera, green (Nikon, #MXR00704) and red (Nikon, #MXR00711) fluorescence filters, as well as bright field. Localization patterns reported in supplementary fig. S9, Supplementary Material online in representative animals were consistently observed in all animals of the same strain ($n = 30$ each for wild-type and ΔHTH GFP::FBXA-215, and $n = 10$ for GFP::FBXA-192). The images in Fig. 3a were created by imaging embryos, which were obtained through bleaching gravid adults, followed by thorough washes in M9 buffer. Afterwards, embryos were introduced into an M9 droplet on a 2% agarose pad and immediately imaged on a Leica SP8 confocal on a DM6000 upright microscope (Leica LAS X image acquisition software, v3.5.7.23225), equipped with an HC PL APO CS2 100×/1.4 oil objective. Imaging was done with sequential scanning with detection ranges maximized for signal. Images shown in Fig. 3a represent a single confocal plane taken with a 1 AU pinhole at 580 nm. Images were processed using Fiji v2.14 (Schindelin et al. 2012), by adjusting brightness and contrast.

Co-localization was quantified in Fiji v2.14 using the JACoP Plugin v2.1.4 (Bolte and Cordelières 2006). Perinuclear regions displaying DEPS-1::RFP signal were isolated from confocal images, from four (GFP::FBXA-215; DEPS-1::RFP) or three (GFP::FBXA-215[ΔHTH]; DEPS-1::RFP) embryos, and these images were used to calculate co-localization coefficients of the signals of DEPS-1::RFP and GFP::FBXA-215. These coefficients were calculated for images corresponding to multiple confocal planes of the same embryo and an average was calculated per strain. Specifically, Pearson's coefficient and Van Steensel's cross-correlation function (CCF) were calculated. Pearson's co-localization coefficient results are indicated in the main text, but all data including CCF analysis are available in supplementary table S6, Supplementary Material online.

RNA Extraction and mRNA Sequencing

N2 wild-type and *fbxa-215(syb4404)* animals grown at 20 °C were synchronized by bleaching and overnight hatching in M9 buffer with gentle mixing. Synchronized L1 animals were brought onto HB101-seeded NGM plates the next day and grown at 20 °C or at 25 °C. Synchronized young adult worms were collected ~36 h (batch grown at 25 °C), or 48 h (batch grown at 20 °C) after plating. Animals were washed off plates with M9, further washed with M9, washed one last time in nuclease-free water, and frozen in dry ice. To collect embryos, synchronized gravid adult worms were bleached ~47 h (batch grown at 25 °C), or 56 h (batch grown at 20 °C) after plating. After two washes in ice-cold M9 and a final wash in nuclease-free water, embryos were frozen in dry ice.

A previously published RNA extraction protocol was used with minor alterations (Almeida et al. 2019). Frozen worm aliquots were thawed, 300 µl of TRIzol (Life Technologies, 15596026) were added and mixed vigorously. Six freeze–thaw cycles were employed to dissolve animals, specifically by freezing tubes in liquid nitrogen for ~30 s, followed by thawing for ~3 min at 37 °C. After each cycle, tubes were mixed vigorously. After the sixth freeze–thaw cycle, samples were spun down at 20,000 × *g* for 3 min and the supernatant was subsequently collected into a fresh tube. Next, 1 volume of 100% ethanol was added to the sample and mixed vigorously. Finally, the mixtures were loaded onto Direct-zol columns (Zymo Research, R2060) and manufacturer's instructions were followed (in-column DNase I treatment was included). Quality control of samples, library preparation (nondirectional, with poly-A enrichment), and mRNA sequencing (Illumina, PE150) was performed by Novogene.

Bioinformatic Analysis

mRNA-Sequencing Analysis of Datasets Generated in Our Study

Illumina adapters and reads with low-quality calls were filtered out using Trimmomatic v0.39 (Bolger et al. 2014) with

options SLIDINGWINDOW:4:28 MINLEN:36. Quality of raw and trimmed fastq files was assessed with fastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and summarized with multiQC v1.11 (Ewels et al. 2016). Gene expression was quantified from trimmed reads using salmon v1.5.1 (Patro et al. 2017), with options --seqBias --gcBias --validateMappings -l A. DESeq2 (Love et al. 2014) and custom scripts (available at https://github.com/migueldvalmeida/F-box_TEs) were used to calculate normalized and TPM counts, generate plots, and conduct statistical tests on an R framework (R Core Team 2021). See R packages used below, in the end of this section.

Trimmed fastq files were mapped to the *C. elegans* genome (WBcel235) using HISAT2 v2.2.1 (Pertea et al. 2016). Resulting SAM files were converted to BAM format and sorted with samtools v1.10 (Li et al. 2009): (i) samtools view -bS; (ii) samtools sort; and (iii) samtools index. To create bigwig files, the BAM alignment files were used as input to bamCoverage v3.5.1, part of the deepTools package (Ramírez et al. 2016), using options --normalizeUsing CPM -of bigwig --binSize 10. Bigwig files of biological replicates were combined using wiggletools (Zerbino et al. 2014) mean and wigToBigWig v4 (Kent et al. 2010). Genome tracks were plotted with custom scripts (available at https://github.com/migueldvalmeida/F-box_TEs) using the Gviz (Hahne and Ivanek 2016) and GenomicFeatures (Lawrence et al. 2013) packages on an R framework (R Core Team 2021).

To quantify TE expression at the TE family level, we first mapped trimmed reads to the *C. elegans* genome (ce10) using STAR v2.5.4b (Dobin et al. 2013) with options --readFiles Command zcat --outSAMtype BAM SortedByCoordinate --outFilterType BySJout --outFilterMultimapNmax 150 --winAnchorMultimapNmax 150 --alignSJoverhangMin 8 --alignSJDBoverhangMin 3 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 10000000 --alignMatesGapMax 100000000. The resulting BAM files were used as inputs for TEtranscripts v2.2.1 (Jin et al. 2015) with options --stranded no --SortByPos. As TEtranscripts input we used ce10 gene and TE annotations. The latter was downloaded from the list of TEtranscripts-compatible annotations created by the Hammell laboratory (available at https://hammelllab.labsites.cshl.edu/software/). DESeq2 (Love et al. 2014) and custom scripts (available at https://github.com/migueldvalmeida/F-box_TEs) were used to calculate normalized counts, generate plots, and conduct statistical tests on an R framework (R Core Team 2021).

For the analysis above, the following R packages were used: tidyverse (Wickham et al. 2019), lattice (Sarkar et al. 2008), eulerr (Larsson et al. 2022), genefilter (Gentleman et al. 2023), pheatmap (Kolde 2019), reshape2 (Wickham 2020), ggrepel (Slowikowski et al. 2023), biomaRt (Durinck et al. 2023), tximport (Love et al. 2023), RColorBrewer (Neuwirth 2022), ashr (Stephens et al. 2023), ggpubr (Kassambara 2023), Genomic Features (Lawrence et al. 2013), patchwork (Pedersen 2023), and UpSetR (Conway et al. 2017).

Analysis of Publicly Available Datasets

Publicly available ChIP-sequencing processed data (Li et al. 2016; Edwards et al. 2021) was used to generate genome tracks. Bedgraph files were downloaded from GEO (accessions GSE81521 and GSE162063) and converted to bigwig files with bedGraphToBigWig v2.8 (Kent et al. 2010). BigWigs were used to generate genome tracks with custom scripts (available at https://github.com/migueldvalmeida/F-box_TEs) using the Gviz (Hahne and Ivanek 2016) and GenomicFeatures (Lawrence et al. 2013) packages on an R framework (R Core Team 2021).

Publicly available raw RNA-sequencing data from *C. elegans* (Almeida et al. 2019; Edwards et al. 2021) was downloaded from public repositories (GEO, accession number GSE162064; SRA, accession number PRJNA497368) and analyzed exactly as described above for the data generated in this study (except for TEtranscripts analysis, which was not conducted). In addition, we used publicly available raw RNA-sequencing data from zebrafish (Collins et al. 2012; White et al. 2017), downloaded from public repositories (BioProject numbers PRJEB2333 and PRJEB12982). These data were analyzed exactly as described above (TEtranscripts analysis was not done), except that reads were mapped to the Ensembl zebrafish genome (Danio rerio GRCz11, or GCA_000002035.4).

Lastly, we used publicly available quantitative (tables of normalized counts) and qualitative (categorization of genes according to tissue-specific expression) data (Ortiz et al. 2014; Serizay et al. 2020).

Reverse transcription quantitative PCR

RNA was extracted from synchronized young adult animal populations as described above. Subsequently, reverse transcription (RT) and quantitative PCR (qPCR) were conducted as previously described (Almeida et al. 2018), with minor changes. RT was performed with 700 ng of RNA using ProtoScript First-Strand cDNA Synthesis Kit (New England Biolabs, #E6300), according to manufacturer's instructions and using an oligo $d(T)_{23}$ VN. After preparation, RT was diluted 1:3 times in nuclease-free water. Afterwards, qPCR reactions were prepared using Power SYBR Green Master Mix (Thermo Fisher Scientific, #4367659) according to manufacturer's instructions. Reactions were prepared for a total of 10 µl, including a primer final concentration of 300 nM, and 1 µl of template RT. A StepOnePlus Real-Time PCR System (Thermo Fischer Scientific) was used to run the qPCR reactions. Cycling conditions were the following: Standard run; 5 min at 95 °C; [40 cycles of 95 °C for 15 s and 60 °C for 45 s]; melt curve calculation [15 s at 95 °C, 1 min at 60 °C and 15 s at 95 °C (1.6 °C/s of increment in temperature)]. Technical duplicates and biological quadruplicates were used. Analysis was performed using the ΔΔCT method (Schmittgen and Livak 2008). pmp-3 was used as a normalization factor (Hoogewijs et al. 2008). Error bars represent the standard deviation of the four biological replicates. Primers used are indicated in supplementary table S5, Supplementary Material online.

Proteome-Wide Structural Interaction Screen

Receiver operating characteristic (ROC) analysis of ColabFold (Jumper et al. 2021; Mirdita et al. 2022) predictions was performed with a library of 255 pairs of true interactors, including entries from *Caenorhabditis elegans*, curated from Vreven et al. (2015) and Chen et al. (2003), CASP15 (https://predictioncenter.org/casp15), PDB (https://www.rcsb.org), and 490 pairs of true noninteractors curated from Blohm et al. (2014). Prediction performance was evaluated by assessing the area under the curve of the ROC curve for

several parameters, with pTM and ipTM selected for downstream thresholding using the Kolmogorov–Smirnov statistic.

The library of preys was compiled based on a previous study (Ortiz et al. 2014), using information of *C. elegans* genes known to be expressed in spermatogenic or oogenic gonads. Entries with duplicated sequences were removed from the proteome/prey library and matched with the sequence of each entry in the FBXA/bait library in preparation for MSA search and 3D model prediction (without Amber relaxation) with ColabFold. Information on pTM, ipTM, and atomic positions was extracted from each prediction to assess interaction potential and for analysis in R (R Core Team 2021). Atoms separated by a distance smaller than the sum of their van der Waals radii were considered clashed and not used for analysis.

Association Between TEs and Gene Families

TEs were annotated in *Caenorhabditis* species listed in supplementary table S4, Supplementary Material online using EDTA v2.0.0 with the option --sensitive 0 (Ou et al. 2019), and unclassified TE copies were filtered out. Gene annotations and PFAM domain annotations were downloaded from WormBase Parasite (Howe et al. 2016, 2017). TE enrichment by PFAM was calculated using a set of custom scripts (available at https://github.com/migueldvalmeida/F-box_TEs). Briefly, for each species, up to 10 classified TE copies closest to every gene ±500 bp were counted, and counts per gene were then summed for each PFAM domain. A thousand random distributions of TE counts per PFAM domain were generated by a bootstrapped shuffling of TE positions across all scaffolds. An empirical cumulative distribution function was used to calculate the distance of the observed number of TEs per PFAM domain from the random distributions, and *P*-values were corrected for multiple comparisons to generate scores.

Localization of *C. elegans* F-box Genes and TEs

Ensembl BioMart (Harrison et al. 2024) was used to collect the genome localizations of all F-box A1, A2, and B genes (identity of these genes can be found in supplementary table S1, Supplementary Material online). A circos plot was created with Circos v0.69-8 (Krzywinski et al. 2009) to visualize F-box gene localization together with TE distribution. TE Density tracks are displayed on the circos plot as the number of features per 100 kb.

Genome tracks with zoomed-in genomic regions on chromosomes II and V were plotted with custom scripts (available at https://github.com/migueldvalmeida/F-box_TEs) using the Gviz (Hahne and Ivanek 2016) and GenomicFeatures (Lawrence et al. 2013) packages on an R framework (R Core Team 2021).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to all members of the Miska and Rödelsperger labs, as well as Ben Luisi, for discussion and suggestions. We would like to thank Helena Santos-Rosa, Honour McCann, and Oliver Weichenrieder for helpful comments. We thank Nicola Lawrence and Richard Butler of the Gurdon Institute Imaging Facility for microscopy training and assistance. We are grateful to Juan Carlos Rueda Silva, David Jordan, and Bryan Tan for technical assistance, and to Marc Ridyard for excellent laboratory management. We are indebted to the Gurdon Institute Media Kitchen for providing reagents and media. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). We acknowledge SunyBiotech for generating some of the worm strains used in this study.

Author Contributions

Conceptualization: M.V.A. and C.R.; Investigation: M.V.A., Z.L., P.R.-G., L.F., J.S., and C.R.; Data Curation: M.V.A., Z.L., P.R.-G., and C.R.; Formal Analysis: M.V.A., Z.L., P.R.-G., A.D., L.F., J.L.P., J.S., X.L., and C.R.; Visualization: M.V.A. and C.R.; Writing—Original Draft: M.V.A. and C.R.; Writing—Review & Editing: all authors contributed; Project Administration: M.V.A., C.R., and E.A.M; Supervision: M.V.A., F.B., C.R., and E.A.M. Resources: F.B., C.R., and E.A.M.

Funding

M.V.A. was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 101027241. A.D. was supported by the Royal Botanic Gardens Kew. P.R.-G. is supported by a Wellcome Trust Investigator Award (222451/Z/21/Z). C.R. is supported by the Max Planck Society. This work was supported by the following grants to E.A.M.: Wellcome Trust Senior Investigator Award (219475/Z/19/Z) and CRUK award (C13474/A27826). The authors also acknowledge core funding to the Gurdon Institute from Wellcome (092096/Z/10/Z, 203144/Z/16/Z) and CRUK (C6946/A24843).

Conflict of Interest

The authors declare that they have no competing interests.

Data Availability

RNA-sequencing data generated in this study have been deposited to GEO under accession no. GSE251877. Proteomics data are available at the ProteomeXchange Consortium via PRIDE under the accession no. PXD048290. All other datasets used in this manuscript are publicly available on WormBase, WormBase ParaSite, Genbank, PDB, GEO, SRA, UniProt, and InterPro.

References

Almeida MV, de Domingues AMJ, Ketting RF. Maternal and zygotic gene regulatory effects of endogenous RNAi pathways. *PLoS Genet*. 2019:15(2):e1007784. https://doi.org/10.1371/journal.pgen.1007784.

Almeida MV, Dietz S, Redl S, Karaulanov E, Hildebrandt A, Renz C, Ulrich HD, König J, Butter F, Ketting RF. GTSF-1 is required for formation of a functional RNA-dependent RNA polymerase complex in *Caenorhabditis elegans*. *EMBO J*. 2018:37(12):e99325. https://doi.org/10.15252/embj.201899325.

Almeida MV, Vernaz G, Putman ALK, Miska EA. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends Genet*. 2022:38(6):529–553. https://doi.org/10.1016/j.tig.2022.02.009.

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev*. 2005:29(2):231–262. https://doi.org/10.1016/j.femsre.2004.12.008.

Athanasouli M, Witte H, Weiler C, Loschko T, Eberhardt G, Sommer RJ, Rödelsperger C. Comparative genomics and community curation further improve gene annotations in the nematode Pristionchus pacificus. *BMC genomics*. 2020:21(1):708.

Babushok DV, Ostertag EM, Kazazian HH. Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol Life Sci*. 2007:64(5):542–554. https://doi.org/10.1007/s00018-006-6453-4.

Bailey AD, Gray LT, Pavelitz T, Newman JC, Horibata K, Tanaka K, Weiner AM. The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells. *DNA Repair (Amst)*. 2012:11(5):488–501. https://doi.org/10.1016/j.dnarep.2012.02.004.

Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. *Nucleic Acids Res*. 2015:43(W1):W39–W49. https://doi.org/10.1093/nar/gkv416.

Barro-Trastoy D, Köhler C. Helitrons: genomic parasites that generate developmental novelties. *Trends Genet*. 2024:40(5):437–448. https://doi.org/10.1016/j.tig.2024.02.002.

Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*. 2014:42(D1):D396–D400. https://doi.org/10.1093/nar/gkt1079.

Böhne A, Zhou Q, Darras A, Schmidt C, Schartl M, Galiana-Arnoux D, Volff J-N. Zisupton—a novel superfamily of DNA transposable elements recently active in fish. *Mol Biol Evol*. 2012:29(2):631–645. https://doi.org/10.1093/molbev/msr208.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014:30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bolte S, Cordelières FP. A guided tour into subcellular colocalization analysis in light microscopy. *J Microsc*. 2006:224(3):213–232. https://doi.org/10.1111/j.1365-2818.2006.01706.x.

Bonchuk A, Balagurov K, Georgiev P. BTB domains: a structural view of evolution, multimerization, and protein–protein interactions. *BioEssays*. 2023:45(2):2200179. https://doi.org/10.1002/bies.202200179.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, *et al*. Ten things you should know about transposable elements. *Genome Biol*. 2018:19(1):199. https://doi.org/10.1186/s13059-018-1577-z.

Brenner S. The genetics of *Caenorhabditis elegans*. *Genetics*. 1974:77(1):71–94. https://doi.org/10.1093/genetics/77.1.71.

Brunquell J, Morris S, Lu Y, Cheng F, Westerheide SD. The genome-wide role of HSF-1 in the regulation of gene expression in *Caenorhabditis elegans*. *BMC Genomics*. 2016:17(1):559. https://doi.org/10.1186/s12864-016-2837-5.

Carelli FN, Cerrato C, Dong Y, Appert A, Dernburg A, Ahringer J. Widespread transposon co-option in the *Caenorhabditis* germline regulatory network. *Sci Adv*. 2022:8(50):eabo4082. https://doi.org/10.1126/sciadv.abo4082.

Chen Q, Bates AM, Hanquier JN, Simpson E, Rusch DB, Podicheti R, Liu Y, Wek RC, Cornett EM, Georgiadis MM. Structural and genome-wide analyses suggest that transposon-derived protein SETMAR alters transcription and splicing. *J Biol Chem*. 2022:298(5):101894. https://doi.org/10.1016/j.jbc.2022.101894.

Chen R, Mintseris J, Janin J, Weng Z. A protein–protein docking benchmark. *Proteins Struct Funct Bioinform*. 2003:52(1):88–91. https://doi.org/10.1002/prot.10390.

Chiang W-C, Ching T-T, Lee HC, Mousigian C, Hsu A-L. HSF-1 regulators DDL-1/2 link insulin-like signaling to heat-shock responses and modulation of longevity. *Cell*. 2012:148(1–2):322–334. https://doi.org/10.1016/j.cell.2011.12.019.

Collins JE, White S, Searle SMJ, Stemple DL. Incorporating RNA-Seq data into the zebrafish Ensembl genebuild. *Genome Res*. 2012:22(10):2067–2078. https://doi.org/10.1101/gr.137901.112.

Conway JR, Lex A, Gehlenborg N. Upsetr: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*. 2017:33(18):mei. https://doi.org/10.1093/bioinformatics/btx364.

Cordaux R, Udit S, Batzer MA, Feschotte C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A*. 2006:103(21):8101–8106. https://doi.org/10.1073/pnas.0601161103.

Coronado-Zamora M, González J. Transposons contribute to the functional diversification of the head, gut, and ovary transcriptomes across Drosophila natural strains. *Genome Res*. 2023:33(9):1541–1553. https://doi.org/10.1101/gr.277565.122.

Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, Feschotte C. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science*. 2021:371(6531):eabc6405. https://doi.org/10.1126/science.abc6405.

Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008:26(12):1367–1372. https://doi.org/10.1038/nbt.1511.

Cutter AD. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. *Mol Biol Evol*. 2008:25(4):778–786. https://doi.org/10.1093/molbev/msn024.

Davis P, Zarowiecki M, Arnaboldi V, Becerra A, Cain S, Chan J, Chen WJ, Cho J, da Veiga Beltrame E, Diamantakis S, *et al*. WormBase in 2022—data, processes, and tools for analyzing *Caenorhabditis elegans*. *Genetics*. 2022:220(4):iyac003. https://doi.org/10.1093/genetics/iyac003.

Dietz S, Almeida MV, Nischwitz E, Schreier J, Viceconte N, Fradera-Sola A, Renz C, Ceron-Noriega A, Ulrich HD, Kappei D, *et al*. The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1. *Nat Commun*. 2021:12(1):2668. https://doi.org/10.1038/s41467-021-22861-2.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics*. 2013:29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.

Durinck S, Huber W, Davis S, Pepin F, Buffalo VS, Smith M. biomaRt: Interface to BioMart databases (i.e. Ensembl). 2023. [accessed 2023 Sep 27]. https://bioconductor.org/packages/biomaRt/.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004:32(5):1792–1797. https://doi.org/10.1093/nar/gkh340.

Edwards SL, Erdenebat P, Morphis AC, Kumar L, Wang L, Chamera T, Georgescu C, Wren JD, Li J. Insulin/IGF-1 signaling and heat stress differentially regulate HSF1 activities in germline development. *Cell Rep*. 2021:36(9):109623. https://doi.org/10.1016/j.celrep.2021.109623.

Eickbush TH, Malik HS. Origins and evolution of retrotransposons. In: *Mobile DNA II*. Washington (DC): ASM; 2007. p. 1111–1144. https://doi.org/10.1128/9781555817954.ch49.

Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016:32(19):3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol Evol*. 2009:1:205–220. https://doi.org/10.1093/gbe/evp023.

Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of mammalian transcription. *Nat Rev Mol Cell Biol*. 2022:23(7):481–497. https://doi.org/10.1038/s41580-022-00457-y.

Garrigues JM, Tsu BV, Daugherty MD, Pasquinelli AE. Diversification of the *Caenorhabditis* heat shock response by Helitron transposable elements. *eLife*. 2019:8:e51139. https://doi.org/10.7554/eLife.51139.

Gentleman R, Carey VJ, Huber W, Hahne F, Taiwo E, Amusat K, Maintainer BP. Genefilter: genefilter: methods for filtering genes from high-throughput experiments. 2023. [accessed 2023 Sep 27]. https://bioconductor.org/packages/genefilter/.

Gilbert W. Why genes in pieces? *Nature*. 1978:271(5645):501–501. https://doi.org/10.1038/271501a0.

Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci*. 2018:27(1):14–25. https://doi.org/10.1002/pro.3235.

Gomez-Pastor R, Burchfiel ET, Thiele DJ. Regulation of heat shock transcription factors and their roles in physiology and disease. *Nat Rev Mol Cell Biol*. 2018:19(1):4–19. https://doi.org/10.1038/nrm.2017.73.

Gozashti L, Hartl DL, Corbett-Detig R. Universal signatures of transposable element compartmentalization across eukaryotic genomes. bioRxiv 2023.10.17.562820. https://doi.org/10.1101/2023.10.17.562820, 18 December 2024, preprint: not peer reviewed.

Gray LT, Fong KK, Pavelitz T, Weiner AM. Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. *PLOS Genet*. 2012:8(9):e1002972. https://doi.org/10.1371/journal.pgen.1002972.

Hahne F, Ivanek R. Visualizing genomic data using gviz and bioconductor. In: Mathé E, Davis S, editors. *Statistical genomics: methods and protocols. Methods in molecular biology*. vol. 1418. New York (NY): Springer; 2016. p. 335–351. https://doi.org/10.1007/978-1-4939-3578-9_16.

Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Becker A, Bennett R, Berry A, Bhai J, *et al*. Ensembl 2024. *Nucleic Acids Res*. 2024:52(D1):D891–D899. https://doi.org/10.1093/nar/gkad1049.

Hoogewijs D, Houthoofd K, Matthijssens F, Vandesompele J, Vanfleteren JR. Selection and validation of a set of reliable reference genes for quantitative sod gene expression analysis in *C. elegans*. *BMC Mol Biol*. 2008:9(1):9. https://doi.org/10.1186/1471-2199-9-9.

Horváth V, Merenciano M, González J. Revisiting the relationship between transposable elements and the eukaryotic stress response. *Trends Genet*. 2017:33(11):832–841. https://doi.org/10.1016/j.tig.2017.08.007.

Howe KL, Bolt BJ, Cain S, Chan J, Chen WJ, Davis P, Done J, Down T, Gao S, Grove C, *et al*. WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res*. 2016:44(D1):D774–D780. https://doi.org/10.1093/nar/gkv1217.

Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite—a comprehensive resource for helminth genomics. *Mol Biochem Parasitol*. 2017:215:2–10. https://doi.org/10.1016/j.molbiopara.2016.11.005.

Huang X, Feng X, Yan Y, Xu D, Wang K, Zhu C, Dong M, HuangX, Guang S, Chen X. Compartmentalized localization of perinuclear proteins within germ granules in *C. elegans*. bioRxiv 2024.03.25.586584. https://doi.org/10.1101/2024.03.25.586584, 28 March 2024, preprint: not peer reviewed.

Iyer LM, Zhang D, de Souza RF, Pukkila PJ, Rao A, Aravind L. Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. *Proc Natl Acad Sci U S A*. 2014:111(5):1676–1683. https://doi.org/10.1073/pnas.1321818111.

Jacob F. Evolution and tinkering. *Science*. 1977:196(4295):1161–1166. https://doi.org/10.1126/science.860134.

Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-Seq datasets. *Bioinformatics*. 2015:31(22):3593–3599. https://doi.org/10.1093/bioinformatics/btv422.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012:337(6096):816–821. https://doi.org/10.1126/science.1225829.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021:596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kaneda Y, Miyata H, Xu Z, Shimada K, Kamoshita M, Nakagawa T, Emori C, Ikawa M. FBXO24 deletion causes abnormal accumulation of membraneless electron-dense granules in sperm flagella and male infertility. *eLife*. 2024:13:RP92794. https://doi.org/10.7554/eLife.92794.3.

Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots. 2023. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/ggpubr/index.html.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013:30(4):772–780. https://doi.org/10.1093/molbev/mst010.

Kawasaki I, Shim Y-H, Kirchner J, Kaminker J, Wood WB, Strome S. PGL-1, a predicted RNA-binding component of germ granules, is essential for fertility in *C. elegans*. *Cell*. 1998:94(5):635–645. https://doi.org/10.1016/S0092-8674(00)81605-0.

Keightley M-C, Carradice DP, Layton JE, Pase L, Bertrand JY, Wittig JG, Dakic A, Badrock AP, Cole NJ, Traver D, *et al*. The pu.1 target gene Zbtb11 regulates neutrophil development through its integrase-like HHCC zinc finger. *Nat Commun*. 2017:8(1):14911. https://doi.org/10.1038/ncomms14911.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*. 2010:26(17):2204–2207. https://doi.org/10.1093/bioinformatics/btq351.

Kim H, Ishidate T, Ghanta KS, Seth M, Conte D Jr, Shirayama M, Mello CC. A co-CRISPR strategy for efficient genome editing in *Caenorhabditis elegans*. *Genetics*. 2014:197(4):1069–1080. https://doi.org/10.1534/genetics.114.166389.

Kipreos ET, Pagano M. The F-box protein family. *Genome Biol*. 2000:1(5):REVIEWS3002. https://doi.org/10.1186/gb-2000-1-5-reviews3002.

Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. . g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res*. 2023:51(W1):W207–W212. https://doi.org/10.1093/nar/gkad347.

Kolde R. pheatmap: Pretty Heatmaps. 2019. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/pheatmap/index.html.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009:19(9):1639–1645. https://doi.org/10.1101/gr.092759.109.

Lanciano S, Mirouze M. Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Curr Opin Genet Dev*. 2018:49:106–114. https://doi.org/10.1016/j.gde.2018.04.002.

Lang D, Chien G. wordcloud2: Create Word Cloud by "htmlwidget". 2018. [accessed 2024 Oct 29]. https://cran.r-project.org/web/packages/wordcloud2/index.html.

Larsson J, Godfrey AJR, Gustafsson P, Privé F. eulerr: area-proportional Euler and Venn diagrams with ellipses. 2022. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/eulerr/index.html.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013:9(8):e1003118. https://doi.org/10.1371/journal.pcbi.1003118.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009:25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li J, Chauve L, Phelps G, Brielmann RM, Morimoto RI. E2f coregulates an essential HSF developmental program that is distinct from the heat-shock response. *Genes Dev*. 2016:30(18):2062–2075. https://doi.org/10.1101/gad.283317.116.

Li Z, Liu X, Zhang Y, Li Y, Zhou L, Yuan S. FBXO24 modulates mRNA alternative splicing and MIWI degradation and is required for normal sperm formation and male fertility. *eLife*. 2024:12:RP91666. https://doi.org/10.7554/eLife.91666.3.

Li Z, Rödelsperger C. Did Caenorhabditis nematodes recycle transposons to fight pathogens? bioRxiv 2022.01.28.478163. https://doi.org/10.1101/2022.01.28.478163, 28 January 2022, preprint: not peer reviewed.

Long M, VanKuren NW, Chen S, Vibranovski MD. New gene evolution: little did we know. *Annu Rev Genet*. 2013:47(1):307–333. https://doi.org/10.1146/annurev-genet-111212-133301.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol*. 2014:15(12):550. https://doi.org/10.1186/s13059-014-0550-8.

Love MI, Soneson C, Robinson M, Patro R, Morgan AP, Thompson RC, Shirley M, Srivastava A. tximport: Import and summarize transcript-level estimates for transcript- and gene-level analysis. 2023. [accessed 2023 Sep 27]. https://bioconductor.org/packages/tximport/.

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022:19(6):679–682. https://doi.org/10.1038/s41592-022-01488-1.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, *et al*. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2021:49(D1):D412–D419. https://doi.org/10.1093/nar/gkaa913.

Morton EA, Lamitina T. *Caenorhabditis elegans* HSF-1 is an essential nuclear protein that forms stress granule-like structures following heat shock. *Aging Cell*. 2013:12(1):112–120. https://doi.org/10.1111/acel.12024.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*. 2012:8(7):e1002764. https://doi.org/10.1371/journal.pgen.1002764.

Neuwirth E. RColorBrewer: ColorBrewer Palettes. 2022. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/RColorBrewer/index.html.

Newman JC, Bailey AD, Fan H-Y, Pavelitz T, Weiner AM. An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in cockayne syndrome. *PLoS Genet*. 2008:4(3):e1000031. https://doi.org/10.1371/journal.pgen.1000031.

Oggenfuss U, Badet T, Croll D. A systematic screen for co-option of transposable elements across the fungal kingdom. *Mob DNA*. 2024:15(1):2. https://doi.org/10.1186/s13100-024-00312-1.

Ortiz MA, Noble D, Sorokin EP, Kimble J. A new dataset of spermatogenic vs. oogenic transcriptomes in the nematode *Caenorhabditis elegans*. *G3 (Bethesda)*. 2014:4(9):1765–1772. https://doi.org/10.1534/g3.114.012351.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019:20(1):275. https://doi.org/10.1186/s13059-019-1905-y.

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet*. 2019:20(2):89–108. https://doi.org/10.1038/s41576-018-0073-3.

Panek J, Gang SS, Reddy KC, Luallen RJ, Fulzele A, Bennett EJ, Troemel ER. A cullin-RING ubiquitin ligase promotes thermotolerance as part of the intracellular pathogen response in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A*. 2020:117(14):7950–7960. https://doi.org/10.1073/pnas.1918417117.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017:14(4):417–419. https://doi.org/10.1038/nmeth.4197.

Patthy L. Genome evolution and the evolution of exon-shuffling—a review. *Gene*. 1999:238(1):103–114. https://doi.org/10.1016/S0378-1119(99)00228-0.

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, *et al*. InterPro in 2022. *Nucleic Acids Res*. 2023:51(D1):D418–D427. https://doi.org/10.1093/nar/gkac993.

Pedersen TL. Patchwork: The composer of plots. 2023. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/patchwork/index.html.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-Seq experiments with HISAT, StringTie and ballgown. *Nat Protoc*. 2016:11(9):1650–1667. https://doi.org/10.1038/nprot.2016.095.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016:44(W1):W160–W165. https://doi.org/10.1093/nar/gkw257.

Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc*. 2007:2(8):1896–1906. https://doi.org/10.1038/nprot.2007.261.

R Core Team. R: A language and environment for statistical computing. 2021. [accessed 2023 Sep 27]. https://www.R-project.org/.

Richardson JM, Colloms SD, Finnegan DJ, Walkinshaw MD. Molecular architecture of the mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell*. 2009:138(6):1096–1108. https://doi.org/10.1016/j.cell.2009.07.012.

Robertson HM, Zumpano KL. Molecular evolution of an ancient mariner transposon, Hsmarl, in the human genome. *Gene*. 1997:205(1-2):203–217. https://doi.org/10.1016/S0378-1119(97)00472-1.

Rockman MV, Kruglyak L. Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet*. 2009:5(3):e1000419. https://doi.org/10.1371/journal.pgen.1000419.

Röseler W, Collenberg M, Yoshida K, Lanz C, Sommer RJ, Rödelsperger C. The improved genome of the nematode Parapristionchus giblindavisi provides insights into lineage-specific gene family evolution. *G3 (Bethesda)*. 2022:12(10):jkac215. https://doi.org/10.1093/g3journal/jkac215.

Saighi P, Jaouadi C, Vieira FRJ, Bernardes JS. DAVI: a tool for clustering and visualising protein domain architectures. bioRxiv 2021.09.24.461671. https://doi.org/10.1101/2021.09.24.461671, 27 September 2021, preprint: not peer reviewed.

Sarkar D, Andrews F, Wright K, Klepeis N. lattice: Trellis Graphics for R. 2008. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/lattice/index.html.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, *et al*. Fiji: an open-source platform for biological-image analysis. *Nat Methods*. 2012:9(7):676–682. https://doi.org/10.1038/nmeth.2019.

Schläpfer P, Mehta D, Ridderikhoff C, Uhrig RG. DomainViz: intuitive visualization of consensus domain distributions across groups of proteins. *Nucleic Acids Res*. 2021:49(W1):W169–W173. https://doi.org/10.1093/nar/gkab391.

Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. *Nat Protoc*. 2008:3(6):1101–1108. https://doi.org/10.1038/nprot.2008.73.

Schreiner WP, Pagliuso DC, Garrigues JM, Chen JS, Aalto AP, Pasquinelli AE. Remodeling of the *Caenorhabditis elegans* noncoding RNA transcriptome by heat shock. *Nucleic Acids Res*. 2019:47(18):9829–9841. https://doi.org/10.1093/nar/gkz693.

Serizay J, Dong Y, Jänes J, Chesney M, Cerrato C, Ahringer J. Distinctive regulatory architectures of germline-active and somatic genes in *C. elegans*. *Genome Res*. 2020:30(12):1752–1765. https://doi.org/10.1101/gr.265934.120.

Sherrill-Mix S. taxonomizr: Functions to work with NCBI accessions and taxonomy. 2023. [accessed 2024 Oct 29]. https://cran.r-project.org/web/packages/taxonomizr/index.html.

Shevchenko A, Tomas H, Havli J, Olsen JV, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc*. 2007:1(6):2856–2860. https://doi.org/10.1038/nprot.2006.468.

Skaar JR, Pagan JK, Pagano M. Mechanisms and function of substrate recruitment by F-box proteins. *Nat Rev Mol Cell Biol*. 2013:14(6):369–381. https://doi.org/10.1038/nrm3582.

Slowikowski K, Schep A, Hughes S, Dang TK, Lukauskas S, Irisson J-O, Kamvar ZN, Ryan T, Christophe D, Hiroaki Y, *et al.* ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2". 2023. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/ggrepel/index.html.

Smythe AB, Holovachov O, Kocot KM. Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. *BMC Evol Biol*. 2019:19(1):121. https://doi.org/10.1186/s12862-019-1444-x.

Snoek BL, Volkers RJM, Nijveen H, Petersen C, Dirksen P, Sterken MG, Nakad R, Riksen JAG, Rosenstiel P, Stastna JJ, *et al.* A multi-parent recombinant inbred line population of *C. elegans* allows identification of novel QTLs for complex life history traits. *BMC Biol*. 2019:17(1):24. https://doi.org/10.1186/s12915-019-0642-8.

Spike CA, Bader J, Reinke V, Strome S. DEPS-1 promotes P-granule assembly and RNA interference in *C. elegans* germ cells. *Development*. 2008:135(5):983–993. https://doi.org/10.1242/dev.015552.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014:30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Stephens M, Carbonetto P, Dai C, Gerard D, Lu M, Sun L, Willwerscheid J, Xiao N, Zeng M. Ashr: Methods for adaptive shrinkage, using empirical Bayes. 2023. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/ashr/index.html.

Stevens L, Félix M-A, Beltran T, Braendle C, Caurcel C, Fausett S, Fitch D, Frézal L, Gosse C, Kaur T, *et al.* Comparative genomics of 10 new Caenorhabditis species. *Evol Lett*. 2019:3(2):217–236. https://doi.org/10.1002/evl3.110.

Suen KM, Braukmann F, Butler R, Bensaddek D, Akay A, Lin C-C, Milonaitytė D, Doshi N, Sapetschnig A, Lamond A, *et al.* DEPS-1 is required for piRNA-dependent silencing and PIWI condensate organisation in *Caenorhabditis elegans*. *Nat Commun*. 2020:11(1):4242. https://doi.org/10.1038/s41467-020-18089-1.

Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006:34(Web Server):W609–W612. https://doi.org/10.1093/nar/gkl315.

Tellier M, Chalmers R. Human SETMAR is a DNA sequence-specific histone-methylase with a broad effect on the transcriptome. *Nucleic Acids Res*. 2019:47(1):122–133. https://doi.org/10.1093/nar/gky937.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023:51(D1):D523–D531. https://doi.org/10.1093/nar/gkac1052.

Thomas JH. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Res*. 2006:16(8):1017–1030. https://doi.org/10.1101/gr.5089806.

Tikanova P, Ross JJ, Hagmüller A, Pühringer F, Pliota P, Krogull D, Stefania V, Hunold M, Koreshova A, Koller A, *et al.* A regulatory module driving the recurrent evolution of irreducible molecular complexes. bioRxiv 2024.09.16.613340. https://doi.org/10.1101/2024.09.16.613340, 18 September 2024, preprint: not peer reviewed.

Vihervaara A, Sistonen L. HSF1 at a glance. *J Cell Sci*. 2014:127(2):261–266. https://doi.org/10.1242/jcs.132605.

Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, *et al.* Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J Mol Biol*. 2015:427(19):3031–3041. https://doi.org/10.1016/j.jmb.2015.07.016.

Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. *Mol Biol Evol*. 2018:35(3):773–777. https://doi.org/10.1093/molbev/msx335.

Wells JN, Feschotte C. A field guide to eukaryotic transposable elements. *Annu Rev Genet*. 2020:54(1):539–561. https://doi.org/10.1146/annurev-genet-040620-022145.

White RJ, Collins JE, Sealy IM, Wali N, Dooley CM, Digby Z, Stemple DL, Murphy DN, Billis K, Hourlier T, *et al.* A high-resolution mRNA expression time course of embryonic development in zebrafish. *eLife*. 2017:6:e30860. https://doi.org/10.7554/eLife.30860.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007:8(12):973–982. https://doi.org/10.1038/nrg2165.

Wickham H. reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package. 2020. [accessed 2023 Sep 27]. https://cran.r-project.org/web/packages/reshape2/index.html.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, *et al.* Welcome to the tidyverse. *J Open Source Softw*. 2019:4(43):1686. https://doi.org/10.21105/joss.01686.

Xie D, Ma Y, Ye P, Liu Y, Ding Q, Huang G, Félix M-A, Cai Z, Zhao Z. A newborn F-box gene blocks gene flow by selectively degrading phosphoglucomutase in species hybrids. *Proc Natl Acad Sci U S A*. 2024:121(46):e2418037121. https://doi.org/10.1073/pnas.2418037121.

Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*. 2017:356(6337):eaaj2239. https://doi.org/10.1126/science.aaj2239.

Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res*. 2005:33(3):946–954. https://doi.org/10.1093/nar/gki236.

Zerbino DR, Johnson N, Juettemann T, Wilder SP, Flicek P. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics*. 2014:30(7):1008–1009. https://doi.org/10.1093/bioinformatics/btt737.

Zevian SC, Yanowitz JL. Methodological considerations for heat shock of the nematode *Caenorhabditis elegans*. *Methods*. 2014:68(3):450–457. https://doi.org/10.1016/j.ymeth.2014.04.015.

Zhang G, Félix M-A, Andersen EC. Transposon-mediated genic rearrangements underlie variation in small RNA pathways. *Sci Adv*. 2024:10(38):eado9461. https://doi.org/10.1126/sciadv.ado9461.