Genome-Wide Characterization, Evolution, Structure, and Expression Analysis of the F-box Genes in *Caenorhabditis*

Ailan Wang (wal2020@126.com)

Northwest Agriculture University: Northwest Agriculture and Forestry University https://orcid.org/0000-0002-1329-017X

Wei Chen

North west A & F University

Shiheng Tao

Northwest Agriculture and Forestry University

Research article

Keywords: F-box gene family, Caenorhabditis, copy number variation, tandem duplication, intron sequence elongation, expression pattern

Posted Date: December 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-119014/v1

Version of Record: A version of this preprint was published at BMC Genomics on December 1st, 2021. See the published version at https://doi.org/10.1186/s12864-021-08189-7.

Abstract

Background: F-box proteins represent a diverse class of adaptor proteins of ubiquition proteasome system (UPS) that play critical roles in cell signaling pathway and immune response. Among closely related organisms of *Caenorhabditis*, tremendous divergence in F-box gene copy numbers was caused by large species-specific expansion and contraction. Although F-box gene number expansion plays an important role in shaping the genomic diversity of *Caenorhabditis*, the mechanisms responsible for the copy number variation of F-box genes and their functional diversification is very poorly understood. In this study, we performed a comprehensive evolution and underlying mechanism analysis of F-box genes in five *Caenorhabditis* species: *C.brenneri*, *C.briggsae*, *C.elegans*, *C.japonica*, *C.remanei*).

Results: Herein, we identified and characterized 594, 192, 377, 39, 1426 F-box homologs in the genome of *C.brenneri*, *C.briggsae*, *C.elegans*, *C.japonica*, *C.remanei* respectively. Our work suggested that extensive species-specific tandem duplication followed by slightly gene loss was the main mechanism responsible for F-box gene number divergence in *Caenorhabditis*. After F-box gene duplication events occurred, several different mechanisms have contributed to gene structural divergence including exon/intron gain/loss, mutation, exonization/pseudoexonization, insertion, deletion, and particularly ubiquitous intron sequence elongation. Based on the analysis of high-throughput RNA sequencing data, we proposed that F-box gene function have diversified by both sub- and neofunctionalization through diverged stage-specific expression patterns in *Caenorhabditis*.

Conclusions: Species-specific tandem duplications as well as trifling gene loss have contributed to the disequilibrium evolution pattern of F-box gene family in *Caenorhabditis*, which lead to complex structural variation as well as diversified functions affecting growth and development within and among *Caenorhabditis.* Taken together, our results provide an overview of F-box genes in *Caenorhabditis* genome and the basis for further functional studies.

Introduction

Novel genes play an important role in the genome evolution of living organisms. Organisms can acquire novel genes through a variety of molecular process, such as genomic rearrangements, retroposition, horizontal gene transfer, de novo origination from non-coding genes, and duplication-divergence of existing genes [1]. The novel genes derived from different evolution mechanisms have distinct molecular signatures, and are not equally active in all genomes. Among all of these evolutionary mechanisms for generating novel genes, gene duplication is a major contributor that facilitates organisms to adapt to dynamically changing environments [2, 3].

Several divergent evolutionary fates have been proposed for duplicated genes [4, 5]. The most likely fate of any duplicated gene is pseudogenization and paralogs is thus either unexpressed or functionless. Given increased gene dosage is beneficial, two gene copies will preserve the original gene function [6], and this such evolutionary process is also referred to as concerted evolution [7]. Another evolutionary fate

is sub-functionalization, in which each daughter gene adopts partial original functions of their parental gene [2]. One of the most important outcomes of gene duplication is new functionalization with one copy undergoing adaptive changes and the other maintaining ancestral function [3, 8, 9]. Each of these processes can be involved in the retention of duplicate genes in different conditions [10–15].

In *Caenorhabditis*, gene duplication has been an important evolutionary force for generating genetic diversity among F-box proteins [16]. F-box proteins are a class of substrate adaptor proteins function in SKP1–CUL1–F-box protein (SCF)-mediated ubiquitination protein degradation pathway [17]. The number of F-box genes varies dramatically among closely related species or subspecies [17]. F-box genes are the largest and fastest evolving gene family in plants [18–21]. For instance, the number of F-box Kelch genes (FBKs) tremendously varies among *Arabidopsis thaliana*, *Oryza sativa*, *Poulus trichocarpa* and *Vitis vinifera* [19]. The large number of F-box proteins in plants might be required by their species-specific metabolism, such as responses to various hormones [22], the ciradian clock and photomorphogenesis [23, 24], flower development [25], and defense responses [26]. However, in most of investigated animals, F-box genes are small in number and quite conserved among closely related species. For instance, F-box gene number varies from 66 to 81 in *Euarchontoglires* [27], and only 42–47 in 12 extant *Drosophila* species [28]. In contrast, F-box genes are massively expanded in *Caenorhabditis*, and the number of F-box genes are even more than one thousand [16]. However, few studies to date have considered how these numerous F-box genes generated in genomes of *Caenorhabditis*, and why they were preserved after generation. To illustrate these intriguing undocumented scientific problems, we investigated the mechanism responsible for F-box gene number variation in *Caenorhabditis*, and their gene structural and functional diversification.

Results

Prediction of F-box genes and their protein domain architectures

F-box genes in five *caenorhabditis* genomes and one outgroup *pristionchus pacificus* were comprehensively identified using a combination approach with HMMER, Psscan, PSI-BLAST and InterProScan. The numbers of identified F-box genes varied greatly from 39 to 1426 in the five species (shown in Table 1), and none but *C.japonica* was less than that of *pristionchus pacificus.* The number of these F-box genes was not in proportion to the total number of genes in respective genome. Intriguingly, the percentage of F-box genes in *C.remanei* was extraordinarily high (up to 4.54%). By contrast, we observed only percentage of 0.13% in *C.japonica* genome.

Table 1
The number of F-box genes identified in six species of nematodes

Species	Hmmer	psscan	Hmmer & psscan	psiblast	F-box genes	Genome genes	Percent
C. brenneri	559	471	437	1	594	30667	1.93%
C.briggsae	181	150	139	0	192	21936	0.88%
C.elegans	356	302	281	0	377	20532	1.84%
C.japonica	35	23	19	0	39	29964	0.13%
C.remanei	1377	1252	1203	1	1426	31444	4.54%
P.pacificus	95	79	77	0	97	29644	0.33%

The numbers of F-box proteins with specific domain architecture for each of these species were shown in Fig. 1. More than half of the F-box proteins in each species shared either FBA2 or FTH domain at C-terminus, and only a small number of F-box genes involved other types of domain. In addition, in a large number of F-box genes of these species, known domain was not found at their C-terminus.

Identification of F-box gene paralogous group and orthologous group

In ENSEMBL database, the homology relationships of genes have been inferred and annotated based on sequence similarity, phylogenetic tree and sequence location on chromosome. Paralogous and orthologous of F-box genes in five species of *Caenorhabditis* were downloaded from ENSEMBL database using biomart respectively. Unexpectedly, F-box domains were lost from a large proportion of the homologous F-box genes. The number of genes with and without F-box domains in each paralogous group and orthologous group were shown in Fig. 2 and Fig. 3 respectively.

Protein sequences from each paragroup and orthogroup were aligned respectively. We found several mechanisms underlying the loss of F-box domains that were present in their homologs: (1) multiple point mutations occurred in F-box domain region; (2) long DNA fragments were inserted into F-box domain regions (3) the whole F-box domain fragments were deleted from the extant gene.

F-box gene number variation and underlying mechanisms

During the evolution of *caenorhabditis*, F-box gene number dynamic variation was inferred by reconciling gene tree and species tree using maximum parsimony method. F-box gene number varied dramatically,

especially after extant *caenorhabditis* speciation (Fig. 4). For instance, F-box gene number has expanded to be 1,426 in *C.remanei*. In contrast, there were no more than two hundred F-box genes in *C.briggsae.* Making clear of the mechanisms responsible for duplicates generation and functional divergence would facilitate understanding the function of these F-box genes.

F-box genes identified in each species were mapped to corresponding chromosomes or contigs depending on sequence assembly by using Ideographica program [29]. As shown in Figure S1, a large number of F-box genes were in tandem arrangement on Chromosoms. The portion of F-box genes generated by tandem duplication were at least 53%, 43%, 52% and 74% for *C.brenneri*, *C.briggsae*, *C.elegans* and *C.remanei* respectively (Table 2).

Table 2
The number of F-box genes generated by tandem duplications

	F-box genes	Single genes	Tandem duplication	Percent
C.brenneri	594	78	315	53%
C.briggsae	192	64	82	43%
C.elegans	377	67	197	52%
C.remanei	1296	96	962	74%

Gene structure divergence among F-box gene paralogs

Unexpectedly, we found that F-box gene paralogs have diverged greatly at genetic structure during their short evolutionary history. Based on the comparison of F-box gene structure and sequence, the following several different mechanisms may contributed to the divergence of those paralogs: 1) sequence exonization/pseudoexonization; 2) sequence insertion and/or deletion in exon; 3) the numerous point mutations that may contribute to functional domain gain or loss; 4) intron sequence elongation (Fig. 5). Detail information of genetic divergence of each analyzed paralogs of *C.elegans* and *C.briggsae* were ideographic in Figure S2 and Figure S3 respectively. Among the mechanisms mentioned above, a noteworthy finding was that intron sequence elongated by serial numerous short sequence repeats in a large number of F-box genes (Fig. 6). Intron sequence elongation widens divergence between paralogs, and also provide evolution materials for genetic reprogramming, novel exon and/or functional domain formation.

Functional divergence of F-box gene duplicates

One mechanism underlying the functional divergence of duplicate genes is diverged temporal- and spatial-specific patterns of gene expression during evolution. Based on RNA-seq data analysis, most members within F-box gene family showed distinct stage-specific high expression pattern (Fig. 7), implying that they may have been neo-functionalization or sub-functionalization via temporal-specific expression pattern in *C.elegans* and *C.briggsae*.

Within one paragroup, several distinct expression pattern were observed, the three paralogs at the bottom of Fig. 7C represented consistent lower expression in each development stage. Although the three paralogs at the top of Fig. 7C still expressed in similar pattern, they were obviously more highly expressed in L4, LE and EE development stages. The remaining seven paralogs, in contrast, have diverged greatly in temporal. Following K-means Cluster and deviation from average method, for all of 42 F-box gene paragroups identified in *C.elegans*, none paragroup was clustered into one group. Similarly, in *C.briggsae*, all of 29 F-box gene paragroups also have diverged in expression pattern. For the closed paralog pairs, only 15 of 99 F-box gene siblings identified in *C.elegans* retained the same expression pattern, and all of 30 F-box gene siblings identified in *C.briggsae* have diverged in stage expression pattern. Thus, we inferred that most of F-box gene duplicates may have diverged in function via different expression pattern.

Discussion

F-box gene identifying approach in Caenorhabditis

In the present study, Hidden Markov model, regular expression, and in combination with InterProScan were used to predict F-box proteins in *Caenorhabditis*. Those proteins that have highly diverged in F-box domain region could not be predicted as F-box proteins although they might still retain F-box protein function. However, more likely, those F-box paralogs that lost F-box domain have evolved into novel functional genes. Although the prediction approach used here was difficult to avoid false negative F-box genes, it was widely applied in numerous studies [30–32]. In addition, identification of F-box genes in human using our approach is highly reliable [27]. Notably, the duplicates of F-box genes identified have diverged substantially at corresponding F-box domain region, which may contribute to their functional divergence. However, this conjecture should be confirmed by experimental evidences in future.

Massive expansion of F-box genes within Caenorhabditis

Although the species *C.brenneri*, *C.briggsae*, *C.elegans* and *C.remanei* of *Caenorhabditis* have diverged at only approximately twenty million years ago [33–35], F-box gene number variation among them were more dramatically than that of *Euarchontoglires* in which species have diverged more than 100 million years [27, 36–38]. Many duplicates of F-box genes rapidly diverged at F-box domain region, such as long

sequence fragment insertion/deletion, and a numerous number of point mutations. Once duplicates emerged, redundant copies may undergo relaxed natural selection pressure, and mutations in sequences provide raw materials for evolution of novel function [4]. Some members of F-box genes were not conserved among *Caenorhabditis*, as the ortholog of the F-box gene in another species was absent. The corresponding ancestral F-box gene may have diverged at F-box domain region which may contribute to the evolution of new traits. The number of F-box genes in *Caenorhabditis* was substantially more than that of other animals [17], and even more than that of plants (Schumann *et al.* 2011). Based on dramatic lineage-specific F-box genes expansion and retention in genomes, we inferred that strongly positive selection may contribute to the fast evolution events in *Caenorhabditis*. F-box genes should be closely linked to special living environment and physiolgoy of *Caenorhabditis.*

F-box genes rapidly diverge at gene structure and function in Caenorhabditis

In the present study, the mechanisms of gene structural and functional divergence of closely related F-box gene paralogs were investigated in *C.briggsae* and *C.elegans*. In such short twenty million years evolutionary history since speciation of *Caenorhabditis* [33], the number of F-box genes massively gain and loss in certain species of *Caenorhabditis*. For instance, *C.elegans* requires F-box protein fog-2 [39] that regulates the translation of tra-2 mRNAs during hermaphrodite development [40]. However, *C.briggsae* lacks fog-2 [41] and instead uses a novel F-box protein *she-1*, that was created by a recent gene duplication and acts upstream of tra-2 as fog-2 does in *C.elegans* [42]. Thus, both species recruited F-box genes produced by recent duplication events into the sex-determination pathway to control hermaphrodite development, but they use distinct paralogs. This result implies not only the number of F-box genes massively gain and loss in certain species of *Caenorhabditis*, but also F-box gene duplicates rapidly diverged at function. In addition, Stage-specific expression pattern of closely related F-box paralogs was widely observed during the bodily development of *C.briggsae* and *C.elegans*, indicating that the function of F-box paralogs may have been new function or sub-function. We conjectured that strong positive selection might drove rapid evolution of F-box genes in *Caenorhabditis*.

F-box genes displayed rapidly gene number variation, structural, functional, and expression pattern divergence, implying that these genes play important function in environmental adaptation and reproduction process [16]. A study showed that SCF complex is involved in response to microsporidiosis and virus mediated by ubiquitin (Bakowski et al. 2014). The target of immune proteasome was ubiquitinated by E3 ubiquitine ligase, although no evidence shows which Culling and adaptor protein were involved in this process. Thomas conjectured that ancestor system of Culling degradation of exogenous proteins is also the ancestor of MHC I [43]. If the conjecture was true, the exongenous and endogenesis adaptor proteins of Culling may be identified by evolution study.

Conclusions

In the present study, we comprehensively analyzed the gene structure, orthologous and paralogous genes, mechanism of gene gain and losses, and gene expression pattern of F-box gene family in five important

Caenorhabditis species. A total of 2725 F-box genes were identified in five species, and 594, 192, 377, 39, 1426 homologs were identified in the genome of *C.brenneri*, *C.briggsae*, *C.elegans*, *C.japonica*, *C.remanei* respectively. In particular, we found that tandem duplications have played an important role in the expansion of the F-box gene family. Mechanisms including exon/intron gain/loss, mutation, exonization/pseudoexonization, insertion, deletion, and particularly ubiquitous intron sequence elongation have contributed to F-box gene structural divergence .Moreover, analyses of their expression profiles provided functional information for members of the F-box gene family in *C.elegans* and *C.briggsae* at different development stages. Importantly, our results shed light on the evolution pattern of F-box genes in *Caenorhabditis* that will provide a valuable resource for future better understanding the biological roles of individual F-box genes.

Methods

Data Retrieval

The proteomic sequences of five *Caenorhabditis* species (*C.brenneri*, *C.briggsae*, *C.elegans*, *C.japonica*, *C.remanei*) and one outgroup species (*P.pacificus*) were downloaded from the ENSEMBL Genome Browser. The Hidden Markov model and Prosite file of F-box domain were downloaded from PFAM (http://pfam.xfam.org/family/f-box#tabview=tab6) [44] and PROSITE respectively (ftp://ftp.expasy.org/databases/prosite/) [45]. Transcriptome sequencing data of different developmental phases of *C.briggsae* and *C.briggsae* was downloaded from modENCODE (http://www.modencode.org/) [46].

Proteome-wide prediction of F-box genes in five species of Caenorhabditis

Hmmersearch program implemented in HMMER software [47] was used to search for F-box domain-containing proteins in five proteome sequences of five *Caenorhabditis* species and *P.pacificus*. We also used regular expression with Perl program ps_scan.pl downloaded from PROSITE [45] to predict F-box genes. Finally, to comprehensively predict F-box proteins that diverged largely at F-box domain, the above identified F-box proteins were used as a PSI-BLAST (e-value = 1e-30) search query against proteome sequences following by confirmation with InterproScan program.

Identification of homology relationship between F-box genes

The paralogs of each F-box gene were downloaded from ENSEMBL using Biomart. Genes that were paralogous to each other were considered as a paralogous group (paragroup). The F-box gene orthologs in five *Caenorhabditis* species were downloaded from ENSEMBL using Biomart. F-box genes that were orthologous to each other were considered as an orthologous group (orthogroup).

Some of genes identified from the above homologous groups search method were not predicted as F-box domain-containing genes. We used sequences alignment to study the mechanism responsible for the

absent of F-box domains from those F-box homologs.

F-box gene number variation and underlying mechanisms

In *Caenorhabditis*, F-box genes are conserved at only F-box domain region. Therefore, full F-box gene sequence is not appropriate for constructing gene trees to infer gene number variation. In the present study, a gene tree was constructed with F-box domain region sequences for each orthologous group. Next, we combined gene tree with the species tree [48, 49] to infer gene number variation using NOTUNG [50]. Finally, we inferred the total variation for all of the F-box genes based on the above mentioned inference method.

The DNA sequences of *C.elegans* and *C.briggsae* have been assembled to whole chromosomes. We used the program Idiographica provided by Toutai Mituyama to map the identified F-box genes to chromosome. Two genes were considered as tandem duplications given there was no more than twenty genes between them [51]. For species that have no assembled chromosomes, we treated a Contig as a chromosome, resulting in underestimates of tandem duplicates.

Divergence of the gene structure of F-box paralogs

A phylogenetic tree was constructed for each identified F-box gene paralogous group. The closest two paralogs were compared for their difference in gene structure. Because of transcriptome sequencing data available for *C.elegans* and *C.briggsae*, the divergence mechanism of F-box gene paralogs in the two species were studied. Each exon sequence was aligned with the sequence of sibling using lfasta program [52]. Next, the similarity between the two compared sequences was shown in graph, the whole process was done with custom perl scripts. A total of 99 and 37 siblings were aligned well in *C.elegans* and *C.briggsae* respectively. The gene structural divergence mechanisms of these paralogs were then researched.

Functional divergence of duplicated F-box genes

F-box genes are a massively expanded gene family, implying that these duplicates may have diverged in function. Thus we studied the mechanism responsible for functional divergence of these identified F-box gene paralogs. RNA-seq technology based transcriptome profiles of *C.elegans* and *C.briggsae* were downloaded from modeENCODE. Genome sequences and GTF files for *C.elegans* and *C.briggsae* were downloaded from ENSEML database for RNA-seq data analysis.

Index files for the two genomes were generated using Bowtie2 [53]. RNA-seq reads were aligned with respective genome by using Tophat software [54], followed by assembling with Cufflinks [55]. Finally, differential expression analysis were performed using Cuffdiff [1]. We referred to program flow in literature [56]. Heatmap ideograph of gene expression difference were drawn with R package gplot from Bioconductor [57]. Development phase specific expression of F-box paralogous group were computed using mean deviation, which was performed in R.

Abbreviations

UPS: ubiquition proteasome system; Caenorhabditis brenneri: *C.brenneri*; Caenorhabditis briggsae *C.briggsae*; Caenorhabditis elegans: *C.elegans*; Caenorhabditis japonica: *C.japonica*; Caenorhabditis remanei: *C.remanei*; SCF: SKP1–CUL1–F-box protein; FBKs: F-box Kelch genes; *P.pacificus*: Pristionchus pacificus; paragroup: paralogous group; orthogroup: orthologous group

Declarations

Acknowledgments

The authors thank to the lab members of the Bioinformatics Center of Northwest A&F University for their valuable advice on research design and paper discussion.

Availability of data and materials

All data generated or analyzed during this study are included in this article and its Additional files.

Authors' contributions

ALW and SHT design the study. WAL and WC performed data curation and analysis. ALW wrote the manuscript, and SHT reviewed and edited the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests

References

1. Cardoso-Moreira M, Long M: **The origin and evolution of new genes**. In: *Evolutionary Genomics.* Springer; 2012: 161-186.
2. Hughes AL: **The evolution of functionally novel proteins after gene duplication**. *Proceedings of the Royal Society of London Series B: Biological Sciences* 1994, **256**(1346):119-124.
3. Bergthorsson U, Andersson DI, Roth JR: **Ohno's dilemma: evolution of new genes under continuous selection**. *Proceedings of the National Academy of Sciences* 2007, **104**(43):17004-17009.
4. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models**. *Nature Reviews Genetics* 2010, **11**(2):97-108.
5. Zhang J: **Evolution by gene duplication: an update**. *Trends in Ecology & Evolution* 2003, **18**(6):292-298.
6. Ohno S: **Evolution by Gene Duplication** *Springer-Verlag Berlin Heidelberg* 1970:160.

7. Nei M, Rooney AP: **Concerted and birth-and-death evolution of multigene families**. *Annu Rev Genet* 2005, **39**:121.
8. Dittmar K, Liberles D: **Evolution after gene duplication**: John Wiley & Sons; 2011.
9. Näsvall J, Sun L, Roth JR, Andersson DI: **Real-time evolution of new genes by innovation, amplification, and divergence**. *Science* 2012, **338**(6105):384-387.
10. Walsh JB: **How often do duplicated genes evolve new functions?***Genetics* 1995, **139**(1):421-428.
11. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**(1):459-473.
12. Walsh B: **Population-genetic models of the fates of duplicate genes**. *Genetica* 2003, **118**(2-3):279-294.
13. Kondrashov FA, Koonin EV: **A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications**. *Trends in Genetics* 2004, **20**(7):287-290.
14. Burki F, Kaessmann H: **Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux**. *Nature genetics* 2004, **36**(10):1061-1063.
15. Innan H: **Population genetic models of duplicated genes**. *Genetica* 2009, **137**(1):19-37.
16. Thomas JH: **Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants**. *Genome Res* 2006, **16**(8):1017-1030.
17. Kipreos ET, Pagano M: **The F-box protein family**. *Genome Biol* 2000, **1**(5):3002.
18. Yang X, Kalluri UC, Jawdy S, Gunter LE, Yin T, Tschaplinski TJ, Weston DJ, Ranjan P, Tuskan GA: **The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants**. *Plant Physiol* 2008, **148**(3):1189-1200.
19. Schumann N, Navarro-Quezada A, Ullrich K, Kuhl C, Quint M: **Molecular evolution and selection patterns of plant F-box proteins with C-terminal kelch repeats**. *Plant Physiol* 2011, **155**(2):835-850.
20. Hua Z, Zou C, Shiu SH, Vierstra RD: **Phylogenetic comparison of F-Box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift**. *PLoS One* 2011, **6**(1):e16219.
21. Navarro-Quezada A, Schumann N, Quint M: **Plant F-Box Protein Evolution Is Determined by Lineage-Specific Timing of Major Gene Family Expansion Waves**. *PloS one* 2013, **8**(7):e68672.
22. Binder BM, Walker JM, Gagne JM, Emborg TJ, Hemmann G, Bleecker AB, Vierstra RD: **The Arabidopsis EIN3 binding F-Box proteins EBF1 and EBF2 have distinct but overlapping roles in ethylene signaling**. *Plant Cell* 2007, **19**(2):509-523.
23. Han L, Mason M, Risseeuw EP, Crosby WL, Somers DE: **Formation of an SCFZTL complex is required for proper regulation of circadian timing**. *Plant J* 2004, **40**(2):291-301.
24. Kim W-Y, Fujiwara S, Suh S-S, Kim J, Kim Y, Han L, David K, Putterill J, Nam HG, Somers DE: **ZEITLUPE is a circadian photoreceptor stabilized by GIGANTEA in blue light**. *Nature* 2007, **449**(7160):356-360.

25. Chae E, Tan QK-G, Hill TA, Irish VF: **An Arabidopsis F-box protein acts as a transcriptional co-factor to regulate floral development**. *Development* 2008, **135**(7):1235-1245.
26. Kim HS, Delaney TP: **Arabidopsis SON1 is an F-box protein that regulates a novel induced defense response independent of both salicylic acid and systemic acquired resistance**. *Sci Signal* 2002, **14**(7):1469.
27. Wang A, Fu M, Jiang X, Mao Y, Li X, Tao S: **Evolution of the F-Box Gene Family in Euarchontoglires: Gene Number Variation and Selection Patterns**. *PLOS ONE* 2014, **9**(4):e94899.
28. Li A, Xu G, Kong H: **Mechanisms underlying copy number variation in F-box genes: evidence from comparison of 12 Drosophila species**. *Biodiversity Science* 2011, **19**(01):3-16.
29. Kin T, Ono Y: **Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat**. *Bioinformatics* 2007, **23**(21):2945-2946.
30. Gupta S, Garg V, Kant C, Bhatia S: **Genome-wide survey and expression analysis of F-box genes in chickpea**. *BMC Genomics* 2015, **16**:67.
31. Jain M, Nijhawan A, Arora R, Agarwal P, Ray S, Sharma P, Kapoor S, Tyagi AK, Khurana JP: **F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress**. *Plant Physiol* 2007, **143**(4):1467-1483.
32. Xu G, Hong M, Nei M, Kong H: **Evolution of F-box genes in plants: Different modes of sequence divergence and their relationships with functional diversification**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(3):835-840.
33. Memar N, Schiemann S, Hennig C, Findeis D, Conradt B, Schnabel R: **Twenty million years of evolution: The embryogenesis of four Caenorhabditis species are indistinguishable despite extensive genome divergence**. *Dev Biol* 2019, **447**(2):182-199.
34. 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.
35. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A *et al*: **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics**. *PLoS Biol* 2003, **1**(2):E45.
36. Betancur RR, Orti G, Pyron RA: **Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes**. *Ecol Lett* 2015, **18**(5):441-450.
37. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J: **Retroposed elements as archives for the evolutionary history of placental mammals**. *PLoS Biol* 2006, **4**(4):e91.
38. Soria-Carrasco V, Castresana J: **Diversification rates and the latitudinal gradient of diversity in mammals**. *Proceedings of the Royal Society B: Biological Sciences* 2012, **279**(1745):4148-4155.
39. Schedl T, Kimble J: **fog-2, a Germ-Line-Specific Sex Determination Gene Required for Hermaphrodite Spermatogenesis in Caenorhabditis elegans**. *Gentics*, **119**:61.

40. Clifford R, Lee1 M-H, Nayak1 S, Ohmachi M, Giorgini F, Schedl T: **FOG-2, a novel F-box containing protein, associates with the GLD-1 RNA binding protein and directs male sex determination in the C. elegans hermaphrodite germline**. *Development*, **127**:5265–5276.
41. Nayak S, Goree J, Schedl T: **fog-2 and the evolution of self-fertile hermaphroditism in Caenorhabditis**. *PLoS Biol* 2005, **3**(1):e6.
42. Guo Y, Lang S, Ellis RE: **Independent Recruitment of F Box Genes to Regulate Hermaphrodite Development during Nematode Evolution**. *Current Biology* 2009, **19**(21):1853-1860.
43. Thomas JH: **Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants**. *Genome Res* 2006, **16**(8):1017-1030.
44. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J: **The Pfam protein families database**. *Nucleic acids research* 2012, **40**(D1):D290-D301.
45. De Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N: **ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins**. *Nucleic Acids Res* 2006, **34**(suppl 2):W362-W365.
46. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project**. *Science* 2010, **330**(6012):1775-1787.
47. Finn R, Clements J: **HMMER3: a new generation of sequence homology search software**. In.; 2010.
48. Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F, Fitch DH: **Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(24):9003-9008.
49. Kiontke KC, Félix M-A, Ailion M, Rockman MV, Braendle C, Pénigault J-B, Fitch DH: **A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits**. *BMC Evolutionary Biology* 2011, **11**(1):339.
50. Chen K, Durand D, Farach-Colton M: **NOTUNG: a program for dating gene duplications and optimizing gene family trees**. *J Comput Biol* 2000, **7**(3-4):429-447.
51. Xu G, Ma H, Nei M, Kong H: **Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification**. *Proc Natl Acad Sci U S A* 2009, **106**(3):835-840.
52. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proc Natl Acad Sci U S A* 1988, **85**(8):2444-2448.
53. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
54. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.
55. Roberts A, Pimentel H, Trapnell C, Pachter L: **Identification of novel transcripts in annotated genomes using RNA-Seq**. *Bioinformatics* 2011, **27**(17):2325-2329.

56. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**(3):562-578.
57. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.

Figures

Domain.arrangement	C.briggsae	C.remanei	C.brenneri	C.elegans	C.japonica	P.pacificus	Sum
F-box FBA2	28	721	188	113	2	0	1052
F-box FTH	60	305	180	173	2	0	720
F-box Others	14	35	33	20	9	11	122
F-box unknown	90	365	193	71	26	86	831
Sum	192	1426	594	377	39	97	2725

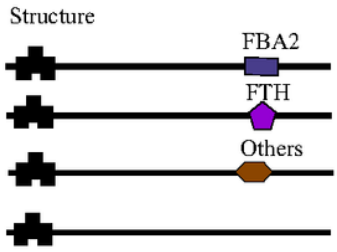

Figure 1

Domain architecture of F-box proteins in six caenorhabditis species

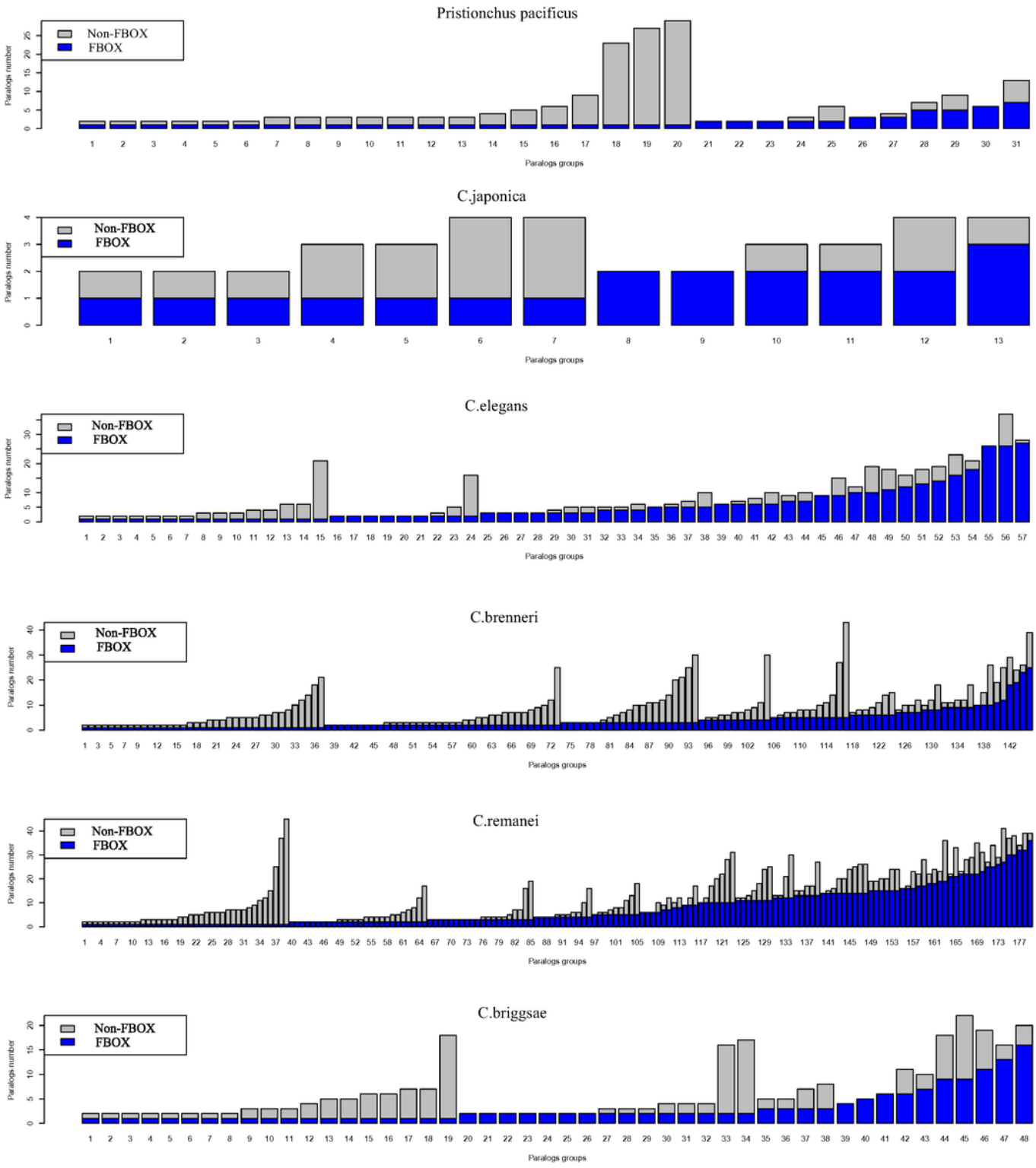

Figure 2

The number of genes with and without F-box domains in each F-box Paragroup from six species Note: Non-FBOX genes denotes genes without F-box domain that present in their homology.

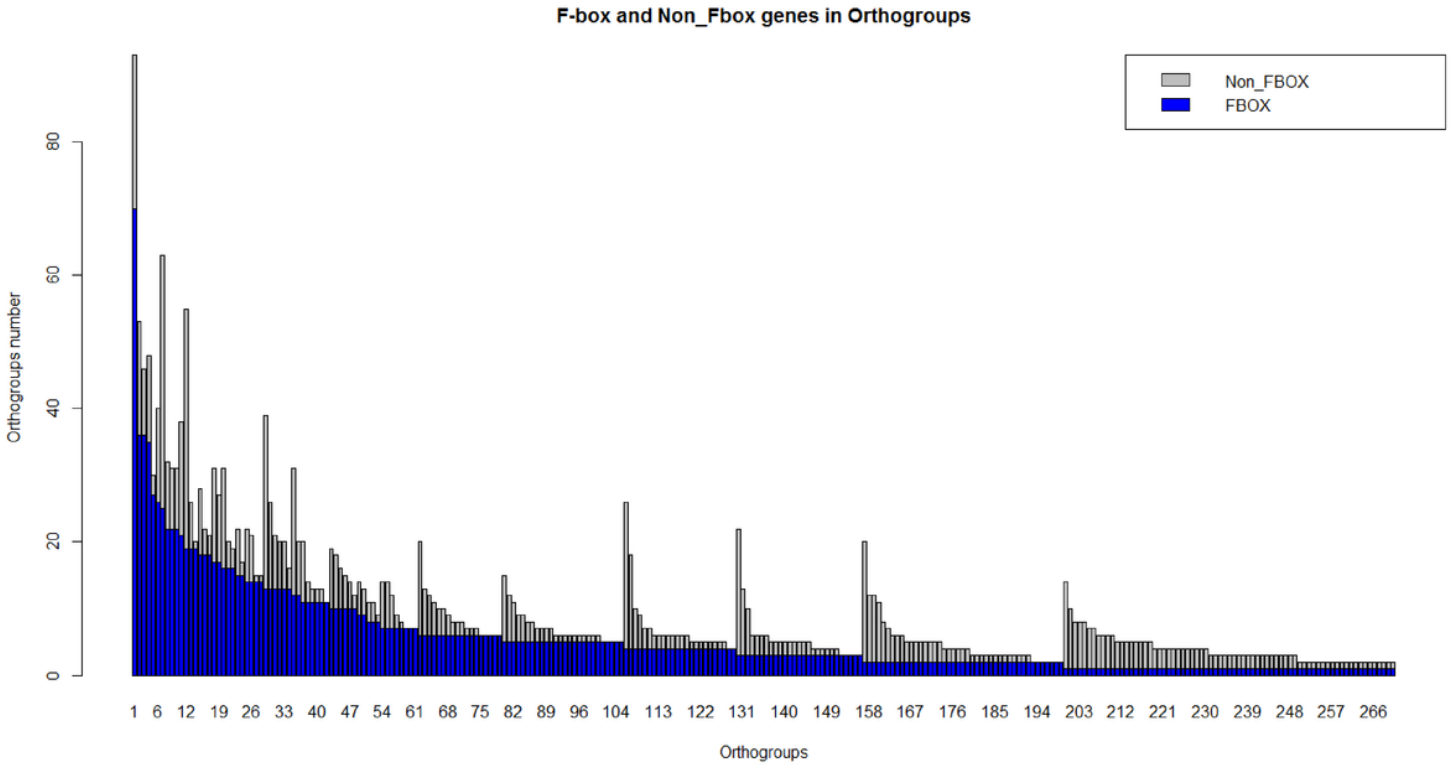

Figure 3

The number of genes with and without F-box domains in each F-box Orthogroup

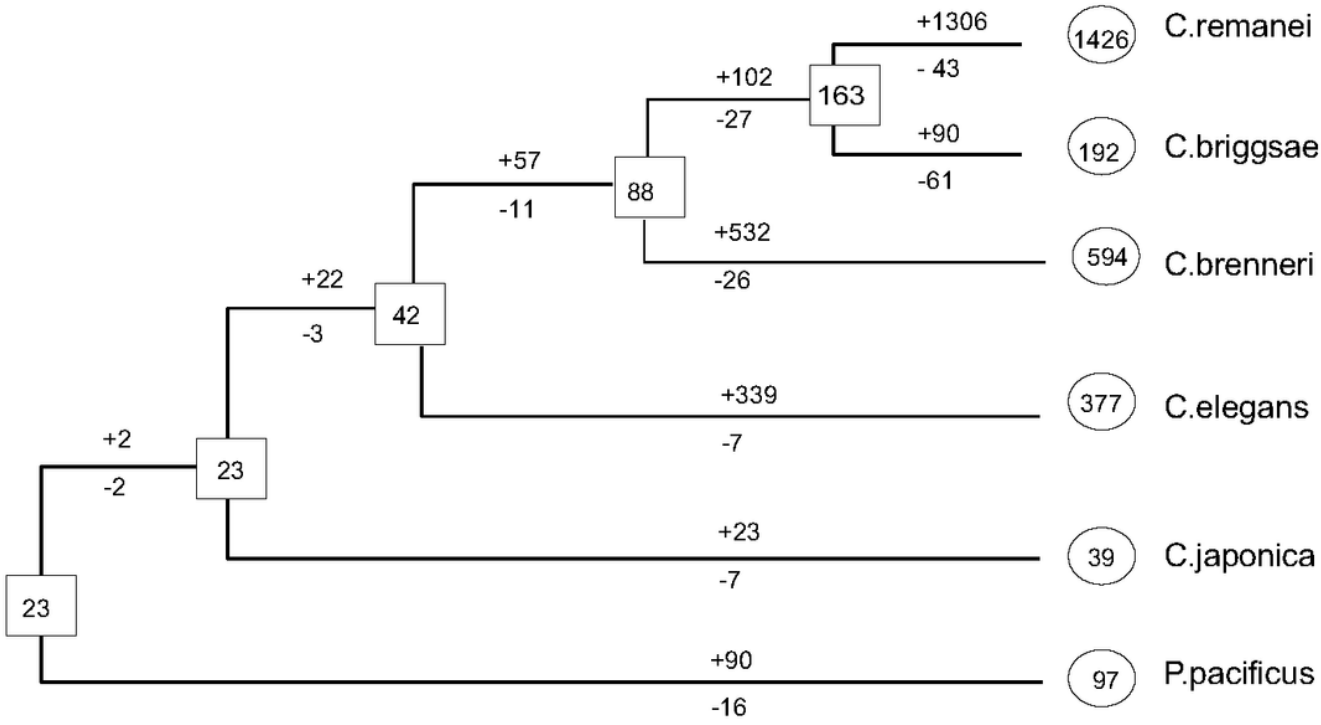

Figure 4

Gain and loss of the number of F-box genes in six nematode species

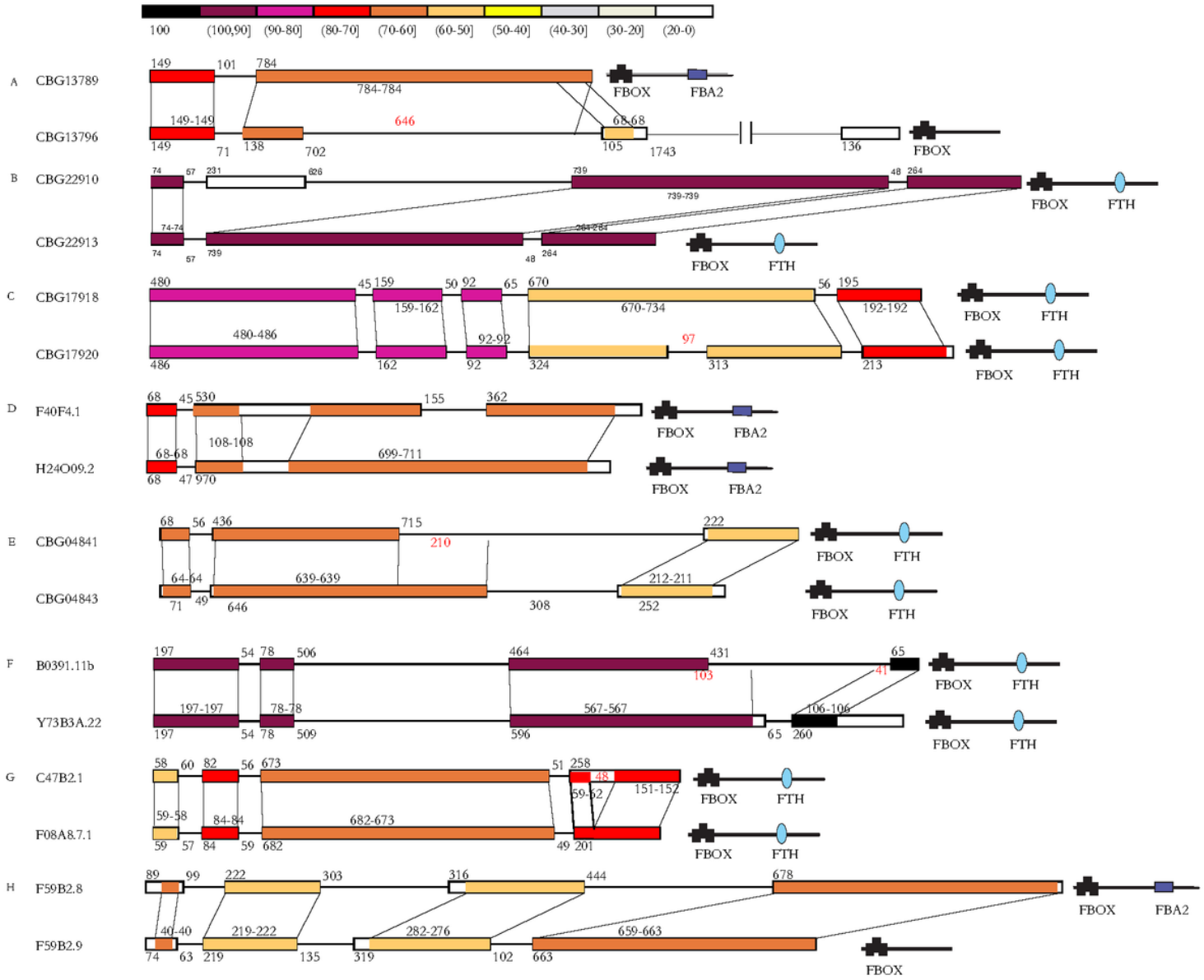

Figure 5

Exon-intron structre of eight pairs of representative sibling paralogs and the domain organization of their proteins.The schematic diagram showed five types of underlying mechanisms responsible for structure divergence of F-box gene paralog. Exons that have experienced exon/intron gain/loss (A-D), exonization/pseudoexonization (A, E, F, G), insertion/deletion (G), point mutations (A, H). The color in schematic diagram represented sequence similarity

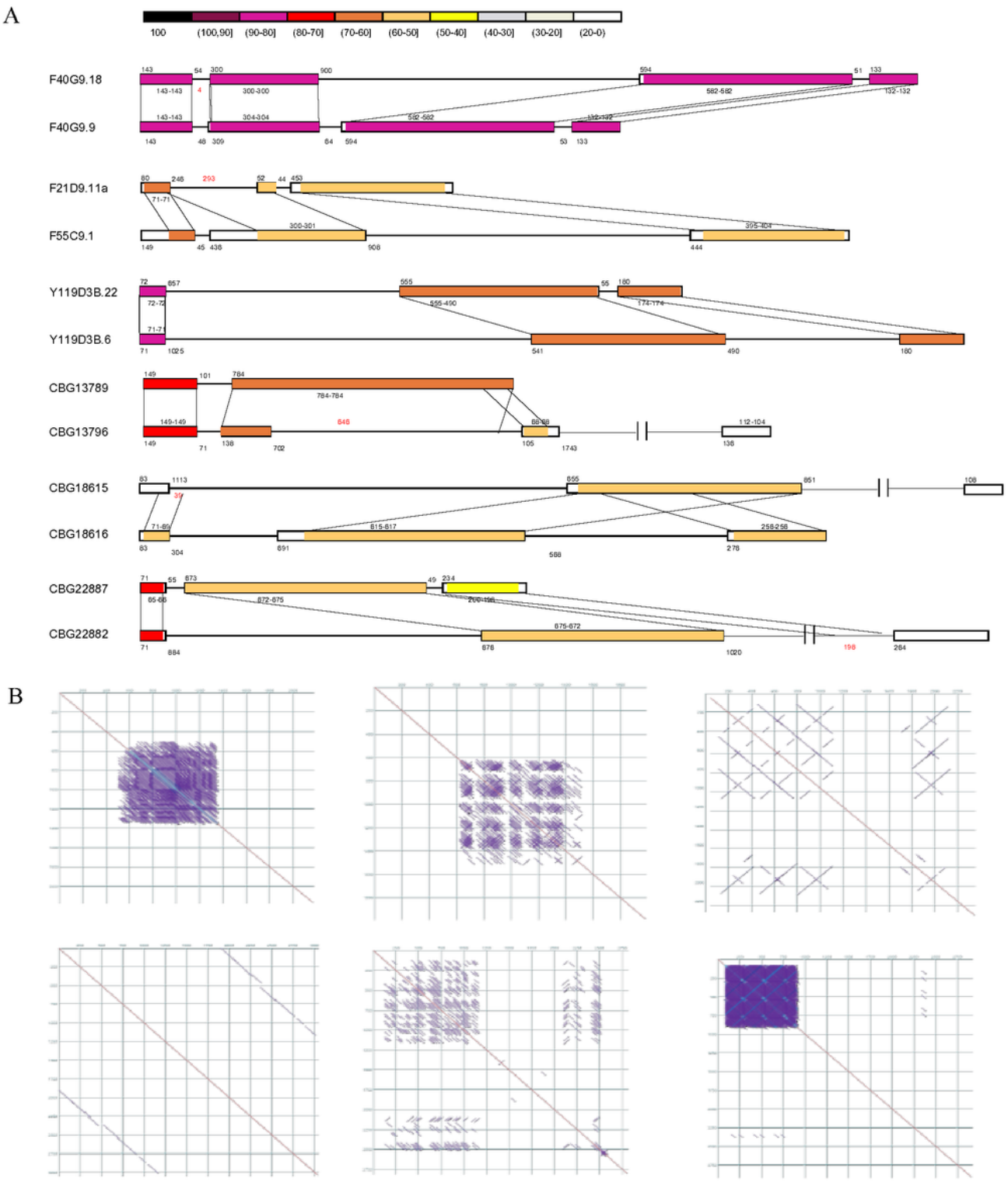

Figure 6

Intron sequence elongation in one paralog of two siblings in six representative pairs. A. Exon-intron structure of six representative pairs of sibling paralogs. B. Gene dotplot of the paralogs with elongated intron sequence generated by numerous short sequence repeats.

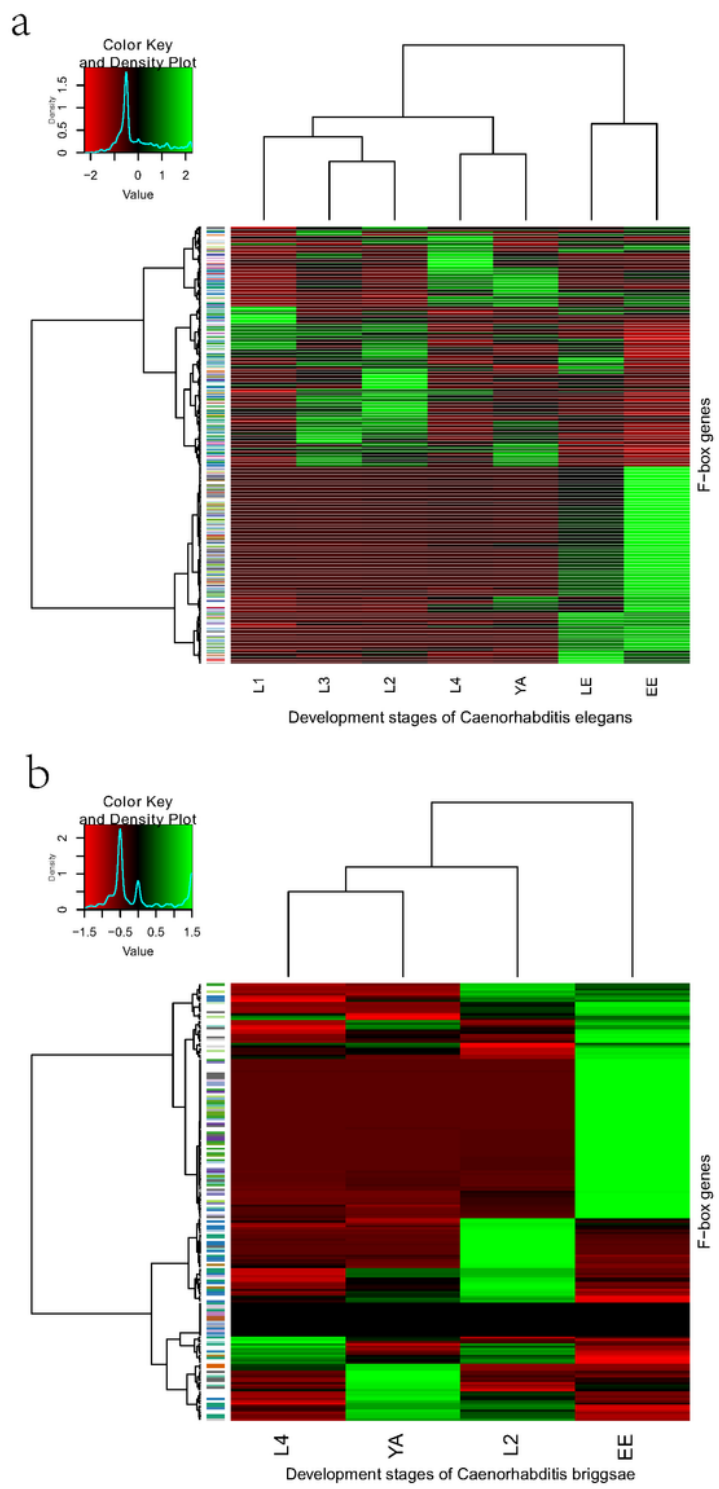

Figure 7

Heatmap of expression pattern divergence of F-box gene duplicates in several development stages of C.elegans and C.briggsae. A. Heatmap of expression profiles of identified F-box genes in seven development stages of C.elegans. B. Heatmap of expression profiles of identified F-box genes in four development stages of C.briggsae. The color bar at the left of heatmap indicates different pararoups (ie, genes in one paragroup with the same color).

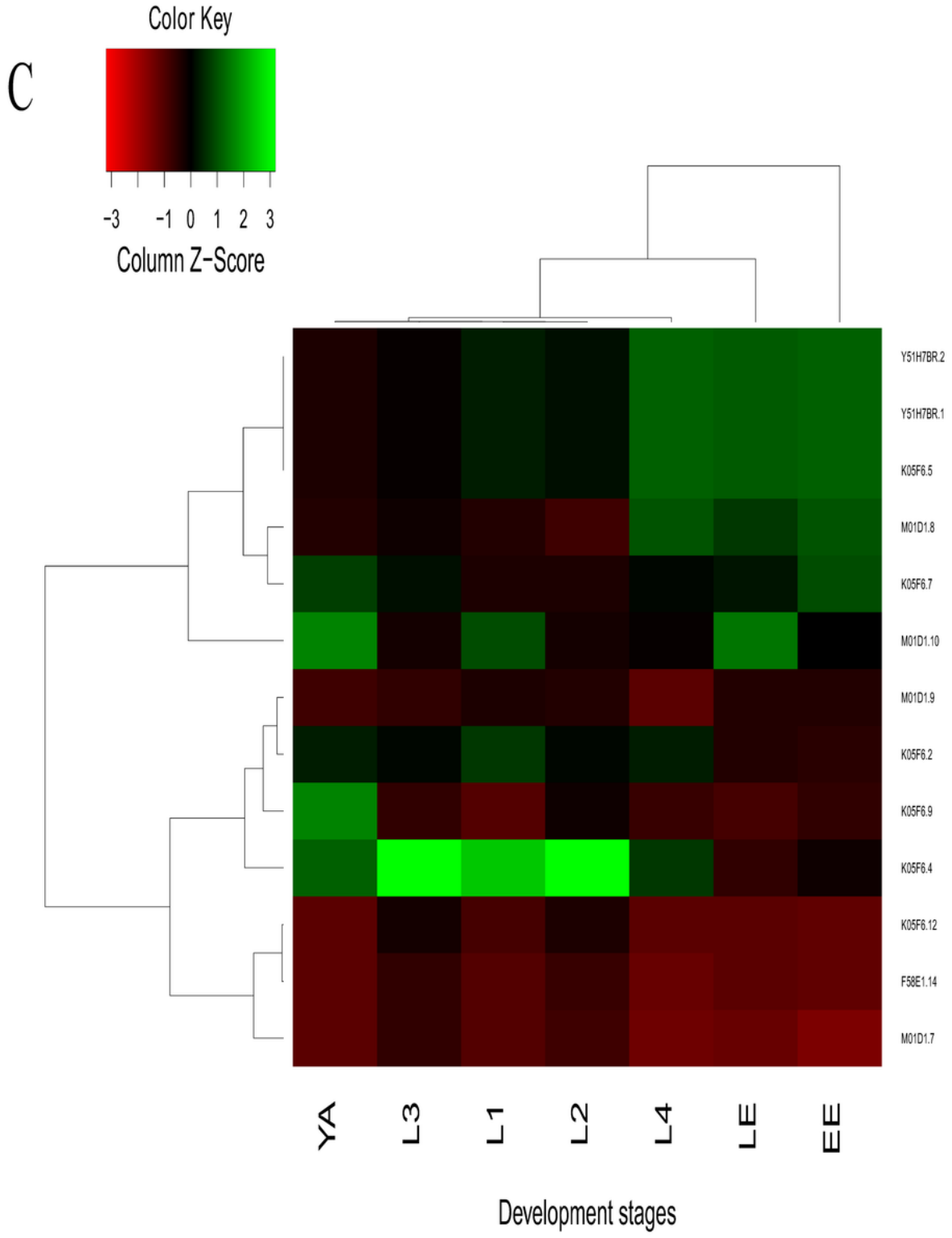

Figure 8

Heatmap of expression difference of a representative F-box gene praragroup in seven development stages of C.elegans, showing discrepant gene expression pattern within paragroup.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1.pdf
- FigS2.pdf
- FigS3.pdf