Worm

Worm

ISSN: (Print) 2162-4054 (Online) Journal homepage: https://www.tandfonline.com/loi/kwrm20

Hitting two birds with one stone: The unforeseen consequences of nested gene knockouts in *Caenorhabditis elegans*

Richard Jovelin & Asher D. Cutter

To cite this article: Richard Jovelin & Asher D. Cutter (2016) Hitting two birds with one stone: The unforeseen consequences of nested gene knockouts in *Caenorhabditis elegans*, Worm, 5:2, e1156835, DOI: 10.1080/21624054.2016.1156835

To link to this article: https://doi.org/10.1080/21624054.2016.1156835

View supplementary material

Published online: 06 Apr 2016.

Submit your article to this journal

Article views: 617

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=kwrm20

WORM
2016, VOL. 5, NO. 2, e1156835 (6 pages)
http://dx.doi.org/10.1080/21624054.2016.1156835

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

Hitting two birds with one stone: The unforeseen consequences of nested gene knockouts in *Caenorhabditis elegans*

Richard Jovelin[a,b] and Asher D. Cutter[a]

[a]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada; [b]Informatics and Bio-Computing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada

ABSTRACT

Nested genes represent an intriguing form of non-random genomic organization in which the boundaries of one gene are fully contained within another, longer host gene. The *C. elegans* genome contains over 10,000 nested genes, 92% of which are ncRNAs, which occur inside 16% of the protein coding gene complement. Host genes are longer than non-host coding genes, owing to their longer and more numerous introns. Indel alleles are available for nearly all of these host genes that simultaneously alter the nested gene, raising the possibility of nested gene disruption contributing to phenotypes that might be attributed to the host gene. Such dual-knockouts could represent a source of misinterpretation about host gene function. Dual-knockouts might also provide a novel source of synthetic phenotypes that reveal the functional effects of ncRNA genes, whereby the host gene disruption acts as a perturbed genetic background to help unmask ncRNA phenotypes.

ARTICLE HISTORY
Received 26 January 2016
Accepted 16 February 2016

KEYWORDS
gene knockout; mutations; nested genes; non-coding RNAs; polymorphism

Introduction

Although many individual labs now routinely perform whole-genome sequencing, understanding how genes interact within genetic pathways and networks remains an important challenge and often represents a limiting factor in genomics. A straightforward way to learn about gene function is to examine the phenotypic effect of ablating a given gene. Forward genetic screens, in which random mutations are selected for a given phenotype, and reverse genetic screens, in which phenotypes are scored for mutations generated in genes of interest, are powerful but laborious approaches. Fifteen years after the completion of the *Caenorhabditis elegans* genome sequence, only ~15% of the protein coding genes have an allele with an associated phenotype.[1] Several large-scale projects provide the worm community with genetic resources to facilitate the investigation of gene function by identifying or generating mutations which can be subsequently introduced in a given genetic background. The *C. elegans* Deletion Mutant Consortium has generated deletion strains for 6,013 genes.[2] The Million Mutation Project has created ~2,000 mutant strains carrying over 800,000 genetic alterations.[3] In addition, whole-genome sequencing of wild isolates provide a rich catalog of naturally occurring allelic variants in *C. elegans* [3,4] and in its relative *C. briggsae*.[5]

Despite these great efforts, the non-random organization of genes in genomes creates practical complications for mapping phenotypes to single genes. Examples of non-random gene organization include clustering of co-expressed genes, a high proportion of genes within operons, and differential gene densities along chromosomes coinciding with variation in recombination rates.[6,7] An especially intriguing and common gene arrangement is when a "nested" gene is located within another "host" gene.[8] Interestingly, in these gene structures, nested and host genes often display weak or negative expression correlation, perhaps because of selection against transcriptional interference.[9-11] Most nested genes are small non-coding RNAs (ncRNAs)[12,13] for which persistence inside host genes appears to be inversely proportional to the ncRNA family size.[14] The most widely studied

CONTACT Richard Jovelin richard.jovelin@oicr.on.ca Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario M5G 0A3, Canada.
Supplemental data for this article can be accessed on the publisher's website.
© 2016 Taylor & Francis

ncRNAs, the microRNAs (miRNAs), are particularly abundant in nested arrangements with ~30% of plant and animal miRNAs located in introns of protein coding genes.[15]

Nested structures pose a challenge for gene knockout studies in which the nested gene may be mistakenly altered along with the focal host gene (or vice versa), making it difficult to ascribe an eventual phenotype to a single gene product. This idea was put into sharp relief by the analysis of gene traps in mouse and the finding that ~200 miRNAs may have concomitantly been misregulated.[16] The functional analysis of miRNAs themselves is complicated by the generation of multiple regulatory small RNAs from a single miRNA gene, and because the different miRNA forms have the potential to bind to the same target genes.[17] Here, we extend the notion that nested genes may be inadvertently disrupted along with their host gene in the worm genome and identify such potential variants. We also identify phenotype-causing alleles for which the coding sequence of host genes has been altered along with the sequence of their nested gene, raising caution for the interpretation of these phenotypes.

Results and discussion

Using the genomic coordinates of 46,734 *C. elegans* genes annotated in WormBase WS248, we identified 10,638 genes that are fully contained in 3,252 host protein coding genes (15.95 % of the 20,391 protein coding genes). Of these nested genes, 9,076 genes are intronic, 636 genes are located within a coding exon and 926 genes are within the boundaries of the host protein coding gene but are neither fully contained within an intron or a coding exon. The majority of host genes (56%) have only 1 nested gene, and up to 146 genes are nested within a host. Host genes tend to be longer than non-host genes and have both more introns and longer introns (Fig. 1), with 19,854 protein coding genes representing candidate hosts by virtue of having at least one intron. Only 608 protein coding genes are nested within a host. In a few instances, the nesting relationships span multiple layers in a way analogous to the matryoshka Russian dolls, with 28 protein coding genes that are both nested and host, and 46 nested genes that are embedded in more than one host gene (Fig. 2). Nested protein coding genes are relatively short, on average 1.6 Kb long, with a median number of 3 introns per gene and a greater proportion of nested genes are intronless (ratio of intronless genes over genes with introns = 0.086) relative to host genes (ratio = 0.003, $\chi^2 = 186.74$, $P < 0.0001$) and non-host genes (ratio = 0.030, $\chi^2 = 46.03$, $P < 0.0001$).

Most of the nested genes (92.25 %) are non-coding RNAs, with piRNAs being the most abundant class (Table 1). The majority of nested genes are located on chromosome IV (60%), with chromosomes I, II and III carrying each 6–7% of nested genes, and with chromosomes V and X having each ~10% of nested gene structures. piRNAs cluster in 2 regions of chromosome IV comprising ~7Mb of sequence,[18] which corresponds to the high density of nested piRNA genes in the genome (Fig. 3). Protein coding genes in the mitochondrial genome are intronless and do not have nested genes.

We then compared the collection of curated variants in Wormbase to gene nestedness. We analyzed 106,511 insertions and deletions (indels) that include lab-generated mutations and natural polymorphisms in wild isolates from the Million Mutation

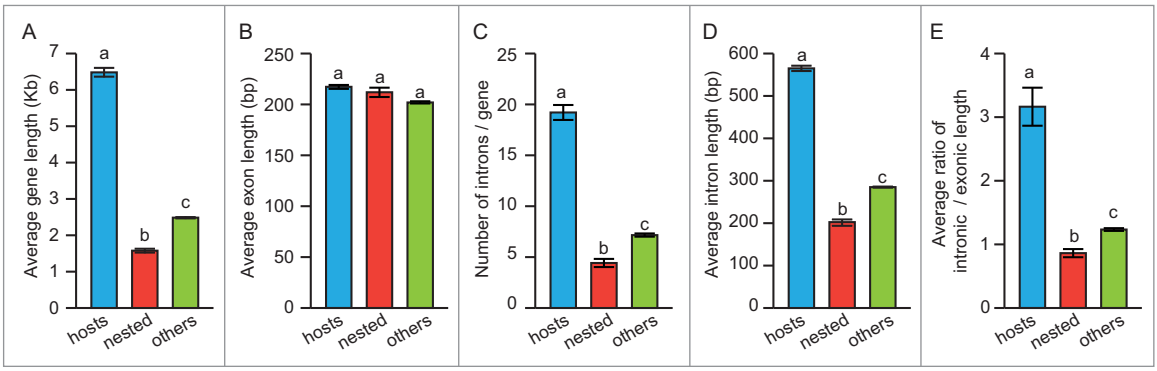

Figure 1. Host genes are longer than protein-coding nested genes and longer then protein-coding genes that are neither host or nested (A). This difference is not due to increased coding exon length in hosts (B) but is due to a higher number of introns (C), longer introns (D) and a greater proportion of intronic sequence over exonic sequence per gene in host genes (E). Means are represented ± 1 standard error. Significant mean differences are indicated by distinct lower capital letters in each panel. *** $P < 0.0001$, Wilcoxon sum rank tests.

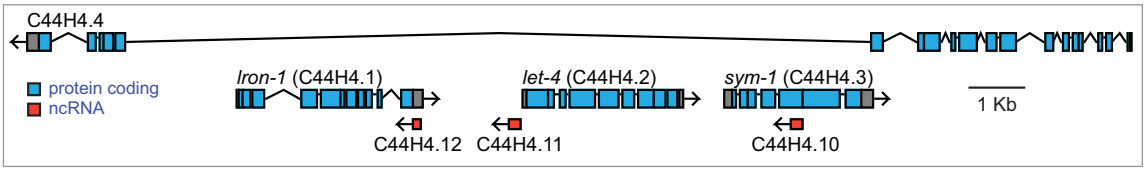

Figure 2. An example of nested gene arrangement in which the nested genes also are host genes.

Project and the Gene Knockout Consortium. We focused on indels because the majority of nested genes are non-coding RNAs located in introns of protein-coding genes, and so single point mutations and single nucleotide polymorphisms (SNPs) are unlikely to influence the functions of both partners in the nested arrangement. In contrast, we hypothesized that indels could alter the expression of the nested genes through complete or partial duplication or deletion. We cross-referenced the positions of host genes and indels and identified 3,400 variants overlapping with the coding sequence of 3,227 host genes and with the sequence of 10,596 nested genes. Thus, there were only 25 host genes and 42 nested genes that lacked variants afflicting both members of the nested-host pair. However, we excluded 2,844 large variants with boundaries falling beyond the positions of the host genes and potentially affecting neighboring genes either directly or indirectly through regulatory sequences, because we are interested in mutations that seemingly alter a focal protein coding gene but that also unintentionally disrupt its nested gene. This procedure yielded a total of 556 variants affecting the coding sequence of 436 host genes along with the sequence of 794 nested genes (Table S1). Consequently, use of these variants to investigate the function of the host protein coding genes presents the risk that any phenotype may be confounded by the alteration of the nested genes. To provide a list of alleles as experimental alternatives, we identified 408 indels that affect only the coding sequence of 199 of these 436 host genes such that they do not simultaneously alter their nested genes (Table S2).

Next we sought to determine how many protein coding genes have an allele with a known phenotype that also disrupts a nested gene. By comparing the list of 3,162 protein coding genes with a phenotype to our list of 436 host genes, we identified 94 alleles with a phenotype that compromise both the coding sequence of 89 host genes and the sequence of their 133 nested genes (Table S3). Again, most nested genes are non-coding RNAs, with the 2 most abundant classes being annotated as ncRNAs and piRNAs (Fig. 4A). Of the 94 alleles, 71 cause a deletion within the host gene and 23 alleles are "complex substitutions" (Fig. 4B), with the median variant lengths being respectively 806 and 672 bp (Fig. 4C) and altering between 1.4% to 100% of the nested gene sequence (Fig. 4D).

As a concrete example of nested gene disruption, consider the gene *unc-59* located on chromosome I. Worms with a 518 bp deletion in *unc-59(tm1928)*, removing most of the second exon, have egg-laying and locomotive defects. This allele also entirely ablates the miRNA *mir-8205* located in the second intron of *unc-59*. In this case, a 329 bp-long deletion (*tm1939*) partially removing *unc-59* exons 1

Table 1. Counts and proportions of nested genes in each functional class.

Functional class	Number of nested genes	Proportion of nested genes (%)	Nesting proportion in each class (%)
piRNA	5881	55.28	38.28
ncRNA	3185	29.94	41.39
protein coding	608	5.72	2.98
tRNA	281	2.64	44.11
Pseudogene	216	2.03	13.29
snoRNA	195	1.83	56.52
miRNA	115	1.08	44.75
snRNA	71	0.67	56.35
asRNA	68	0.64	68.00
rRNA	12	0.11	54.55
lincRNA	5	0.05	2.96
scRNA	1	0.01	100
Total	10638	100	

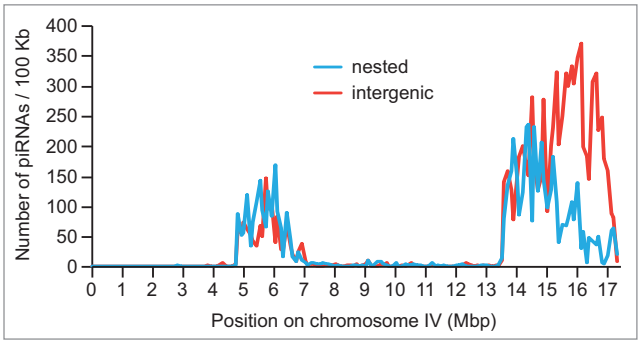

Figure 3. Nested and intergenic piRNAs correspond to the 2 piRNA clusters on chromosome IV. The number of piRNAs is plotted based on 100 Kb long windows.

and 2 has a similar phenotype to the deletion that also removes *mir-8205*, giving confidence that the *tm1928* phenotype is not driven solely by the knockout of *mir-8205*. However, miRNA mutants generally display subtle phenotypes, if at all, with additional environmental or genetic perturbations often facilitating the expression of their functional effects.[19-23] Consequently, might the joint disruption of a nested miRNA and its host gene yield the genetically perturbed conditions that could assist in producing synthetic miRNA phenotypes? In this example, it is unclear how or whether the *mir-8205* deletion could interact directly with *unc-59* to produce additional phenotypic consequences. As *mir-8205* could regulate as many as 162 target genes, its deletion potentially misregulates many transcripts because its unique seed sequence predicts little redundancy with other miRNAs (miRBase.org). More generally, however, it may well be that the disruption of the nested genes unpredictably interfere with the host gene's phenotype through shared genetic network architecture, perhaps by reinforcing or contributing to it in ways unrelated to the host gene's primary function.

In conclusion, we concur with Osokine et al.[16] that the complexity of genome organization, and nested gene structures in particular, increases the risk that the causality of gene function may sometimes be mistakenly interpreted. A further extension of this phenomenon is the possibility of inadvertent disruption of downstream operon gene expression owing to knockout of an upstream gene within a given operon. This may be particularly relevant in worm genetics. As next-generation sequencing costs continue to drop, there may be a renewed interest in forward

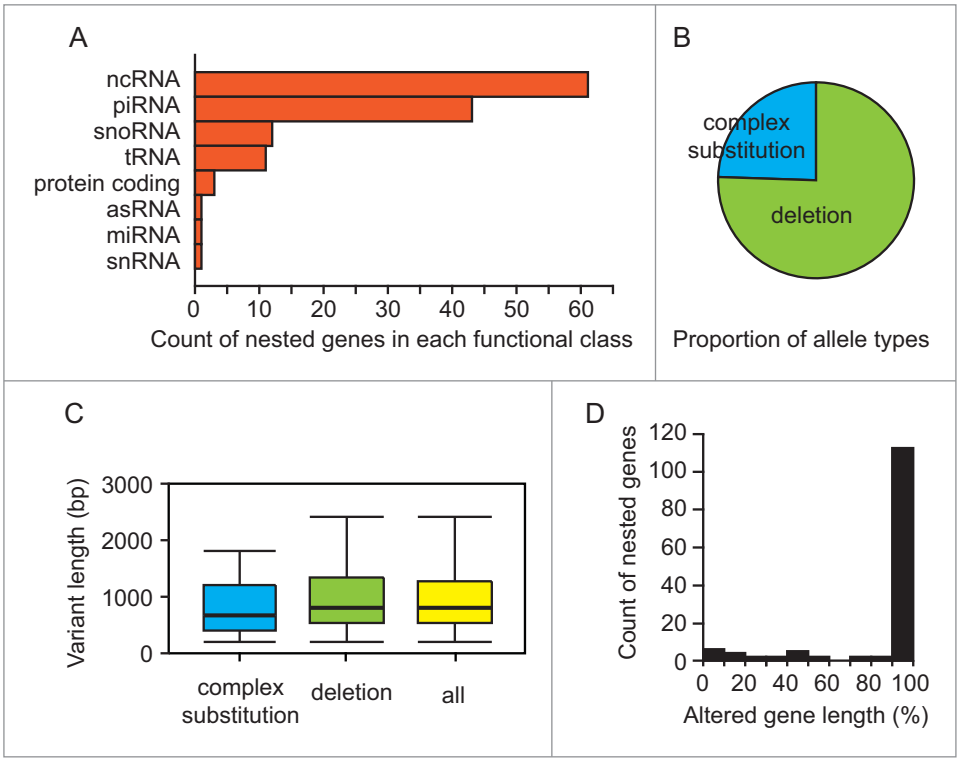

Figure 4. Characteristics of variants affecting both the coding sequence of host genes and the sequence of their nested genes. A. Functional distribution of the nested genes. B. Proportions of the types of variants. C. Median variant length. D. Distribution of the proportion of nested gene length altered by the variants.

genetic screens coupled with whole-genome sequencing to dissect gene function instead of relying on the faster and large-scale RNAi screens.[24,25] Fortunately, most nested genes are non-coding RNAs, for which deletion often results in little developmental defects, and the *C. elegans* genome is exceptionally well-annotated.[26] Consequently, the risk of misinterpreting the genetic basis of knockout phenotypes may be limited to conditions in which the disruption of the ncRNAs has potent effects. On the other hand, an unexpected consequence of mutations that disrupt both a host and its nested genes may be to increase the likelihood of observing a phenotype from the joint knockout. Such synthetic mutants might reveal interesting biological processes, even if potentially more difficult to interpret than single gene knockouts.

Material and methods

We extracted the genomic positions of all 46,734 protein-coding and non-coding genes using the genome annotation of *C. elegans* WS248. We also extracted the genomic coordinates of 106,511 insertion/deletion (indels) variants using the same GFF annotation file. We first searched the *C. elegans* genome for genes that are fully contained within the boundaries of a protein coding gene using the genes' genomic coordinates. We then sorted the nested genes in 3 categories: genes entirely nested within an intron, genes entirely contained within a coding exon, and genes that are within the boundaries of the host but are neither fully intronic or exonic. We then generated a list of variants potentially altering the function of both genes in each host-nested gene pairs by identifying indels with genomic coordinates overlaping with both the coding sequence of the host gene and with the sequence of the nested gene. Using WormMine WS238, we downloaded the list of alleles and their associated phenotypes for 3,162 genes. We then cross-referenced our list of variants with this list of alleles to identify host genes with phenotypes that potentially result from the joint disruption of the nested gene and the disruption of the host's coding sequence. We used TargetScan [27] to predict the target genes of *mir-8205*. We first extracted the sequences of the annotated 3′-untranslated regions (UTRs) of the *C. elegans* genes, keeping the UTR of a single transcript, and we predicted targets using the seed sequences of the 5′-arm and 3′-arm of *mir-8205*.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This work is supported by a grant from the National Health Institutes (GM096008) to A.D.C.

References

[1] Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res 2010; 38:D463-7; PMID: 19910365; http://dx.doi.org/10.1093/nar/gkp952

[2] Consortium TDM. Large-scale screening for targeted knockouts in the *Caenorhabditis elegans* genome. G3 (Bethesda) 2012; 2:1415-25; PMID:23173093; http://dx.doi.org/full_text

[3] Thompson O, Edgley M, Strasbourger P, Flibotte S, Ewing B, Adair R, Au V, Chaudhry I, Fernando L, Hutter H, et al. The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. Genome Res 2013; 23:1749-62; PMID:23800452; http://dx.doi.org/10.1101/gr.157651.113

[4] Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Felix MA, Kruglyak L. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat Genet 2012; 44:285-90; PMID:22286215; http://dx.doi.org/10.1038/ng.1050

[5] Thomas CG, Wang W, Jovelin R, Ghosh R, Lomasko T, Trinh Q, Kruglyak L, Stein LD, Cutter AD. Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. Genome Res 2015; 25:667-78; PMID:25783854

[6] Jovelin R, Dey A, Cutter AD. Fifteen years of evolutionary genomics in *Caenorhabditis elegans*. In: eLS. John Wiley & Sons, Ltd: Chichester, 2013. http://dx.doi.org/10.1002/9780470015902.a0022897.

[7] Cutter AD, Dey A, Murray RL. Evolution of the *Caenorhabditis elegans* genome. Mol Biol Evol 2009; 26:1199-234; PMID:19289596; http://dx.doi.org/10.1093/molbev/msp048

[8] Assis R, Kondrashov AS, Koonin EV, Kondrashov FA. Nested genes and increasing organizational complexity of metazoan genomes. Trends Genet 2008; 24:475-8; PMID:18774620; http://dx.doi.org/10.1016/j.tig.2008.08.003

[9] Yu P, Ma D, Xu M. Nested genes in the human genome. Genomics 2005; 86:414-22; PMID:16084061; http://dx.doi.org/10.1016/j.ygeno.2005.06.008

[10] Lee YC, Chang HH. The evolution and functional significance of nested gene structures in *Drosophila melanogaster*. Genome Biol Evol 2013; 5:1978-85; PMID:24084778; http://dx.doi.org/10.1093/gbe/evt149

[11] Chen N, Stein LD. Conservation and functional significance of gene topology in the genome of *Caenorhabditis elegans*. Genome Res 2006; 16:606-17; PMID:16606698; http://dx.doi.org/10.1101/gr.4515306

[12] St Laurent G, Shtokalo D, Tackett MR, Yang Z, Eremina T, Wahlestedt C, Urcuqui-Inchima S, Seilheimer B, McCaffrey TA, Kapranov P. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. BMC Genomics 2012; 13:504; PMID:23006825; http://dx.doi.org/10.1186/1471-2164-13-504

[13] Mattick JS, Makunin IV. Small regulatory RNAs in mammals. Hum Mol Genet 2005; 14 Spec No 1:R121-32; http://dx.doi.org/10.1093/hmg/ddi101

[14] Wang PP, Ruvinsky I. Family size and turnover rates among several classes of small non-protein-coding RNA genes in *Caenorhabditis* nematodes. Genome Biol Evol 2012; 4:565-74; PMID:22467905; http://dx.doi.org/10.1093/gbe/evs034

[15] Axtell MJ, Westholm JO, Lai EC. Vive la difference: biogenesis and evolution of microRNAs in plants and animals. Genome Biol 2011; 12:221; PMID:21554756; http://dx.doi.org/10.1186/gb-2011-12-4-221

[16] Osokine I, Hsu R, Loeb GB, McManus MT. Unintentional miRNA ablation is a risk factor in gene knockout studies: a short report. PLoS Genet 2008; 4:e34; PMID:18282110

[17] Chen CZ. An unsolved mystery: the target-recognizing RNA species of microRNA genes. Biochimie 2013; 95:1663-76; PMID:23685275; http://dx.doi.org/10.1016/j.biochi.2013.05.002

[18] Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell 2006; 127:1193-207; PMID:17174894; http://dx.doi.org/10.1016/j.cell.2006.10.040

[19] Miska EA, Alvarez-Saavedra E, Abbott AL, Lau NC, Hellman AB, McGonagle SM, Bartel DP, Ambros VR, Horvitz HR. Most *Caenorhabditis elegans* microRNAs are individually not essential for development or viability. PLoS Genet 2007; 3:e215; PMID:18085825; http://dx.doi.org/10.1371/journal.pgen.0030215

[20] Li X, Cassidy JJ, Reinke CA, Fischboeck S, Carthew RW. A microRNA imparts robustness against environmental fluctuation during development. Cell 2009; 137:273-82; PMID:19379693; http://dx.doi.org/10.1016/j.cell.2009.01.058

[21] Alvarez-Saavedra E, Horvitz HR. Many families of *C. elegans* microRNAs are not essential for development or viability. Curr Biol 2010; 20:367-73; PMID:20096582; http://dx.doi.org/10.1016/j.cub.2009.12.051

[22] Brenner JL, Jasiewicz KL, Fahley AF, Kemp BJ, Abbott AL. Loss of individual microRNAs causes mutant phenotypes in sensitized genetic backgrounds in *C. elegans*. Curr Biol 2010; 20:1321-5; PMID:20579881; http://dx.doi.org/10.1016/j.cub.2010.05.062

[23] Park CY, Jeker LT, Carver-Moore K, Oh A, Liu HJ, Cameron R, Richards H, Li Z, Adler D, Yoshinaga Y, et al. A resource for the conditional ablation of microRNAs in the mouse. Cell Rep 2012; 1:385-91; PMID:22570807; http://dx.doi.org/10.1016/j.celrep.2012.02.008

[24] Zuryn S, Jarriault S. Deep sequencing strategies for mapping and identifying mutations from genetic screens. Worm 2013; 2:e25081; PMID:24778934; http://dx.doi.org/10.4161/worm.25081

[25] Bowerman B. The near demise and subsequent revival of classical genetics for investigating *Caenorhabditis elegans* embryogenesis: RNAi meets next-generation DNA sequencing. Mol Biol Cell 2011; 22:3556-8; PMID:21960050; http://dx.doi.org/10.1091/mbc.E11-03-0185

[26] Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 2010; 330:1775-87; PMID:21177976; http://dx.doi.org/10.1126/science.1196914

[27] Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. Nature 2011; 469:97-101; PMID:21085120; http://dx.doi.org/10.1038/nature09616