Chapter 2

Fundamentals of Comparative Genome Analysis in *Caenorhabditis* Nematodes

Eric S. Haag and Cristel G. Thomas

Abstract

The genome of the nematode *Caenorhabditis elegans* was the first of any animal to be sequenced completely, and it remains the "gold standard" for completeness and annotations. Even before the *C. elegans* genome was completed, however, biologists began examining the generality of its features in the genomes of other *Caenorhabditis* species. With many such genomes now sequenced and available via WormBase, *C. elegans* researchers are often confronted with how to interpret comparative genomic data. In this article, we present practical approaches to addressing several common issues, including possible sources of error in homology annotations, the often complex relationships between sequence similarity, orthology, paralogy, and gene family evolution, the impact of sexual mode on genome assemblies and content, and the determination and use of synteny as a tool.

Key words Comparative genomics, Phylogeny, Homology, Paralog, Ortholog

1 Introduction

The sequencing of the *C. elegans* genome in 1998 [1] was a landmark in modern biology. For experimental biologists, it became possible to perform reverse-genetic experiments genome-wide [2–5], create reporter transgene constructs [6], and quantify expression of all genes simultaneously (e.g. [7–10]). For workers in other systems, it inspired the development of genomic methods. Most relevant here, however, it also served as a beachhead from which to launch comparative studies with close relatives of *C. elegans* [11]. Indeed, large-scale characterization of the genomes of non-*elegans* species of *Caenorhabditis* began well before 1998. In 1981, Butler et al. found abundant differences in electrophoretic mobility of homologous enzymes from *C. elegans* and *C. briggsae* [12], implying a surprising amount of molecular divergence in these anatomically very similar animals. An early study of vitellogenin homologs [13] confirmed that homologous sequences between these species were indeed highly divergent, yet also noted

David Biron and Gal Haspel (eds.), *C. elegans: Methods and Applications*, Methods in Molecular Biology, vol. 1327,
DOI 10.1007/978-1-4939-2842-2_2,

that remaining conservation is likely to represent important functional constraints. In the mid-1990s the Genome Sequencing Center at Washington University in St. Louis generated 12 Mbp of *C. briggsae* sequence from large-insert clones. This relatively small amount of sequence was immediately used by many authors, and a few years later an entire assembly was produced by Waterston and his colleagues [14]. As whole-genome shotgun methods were improved and their costs reduced, many new *Caenorhabditis* species were added to the collection, most quickly populating WormBase with varying degrees of curation (http://www.wormbase.org). More recently, multispecies whole-genome comparisons have shed light on subjects as diverse as discovery of noncoding RNAs [15], the introduction of insertion–deletion mutations [16], and the evolution of sexually dimorphic gene expression [17].

In 2015, a user of WormBase examining a *C. elegans* gene page will see a number of fields related to homology. These include both named (e.g. "*Cre-unc-119*") and unnamed (e.g. CBG12344) homologs from other *Caenorhabditis* nematodes, as well as from much more distant relatives, such as *Pristionchus*, parasitic nematodes from different families, mouse, *Drosophila*, human, and various fish. This information puts relationships of "your favorite gene" a mouse-click away, and has greatly increased the speed with which homologs can be found. Yet at the same time, most determinations of homology are based on automated annotation pipelines, and are subject to errors of various sorts. Here, we summarize the major sources of imprecision in homology assignments, present a simple dichotomous key to guide a researcher to a well-supported model of gene evolution, and discuss some computational approaches to working with large datasets.

2 *Caveat Emptor*: Sources of Ambiguity in Homology Assignment

The genome assembly for *C. elegans* is of exceptional quality, but this is not generally true for most other *Caenorhabditis* species. There are two reasons for this discrepancy, as discussed below in Subheadings 2.1 and 2.2.

2.1 Assembly Approach

C. elegans was meticulously assembled using a minimal tiling path of large-insert clones [1], an approach generally not taken in subsequent projects. There was also a wealth of genetic data supporting the linkage of particular sequences on the same chromosome. These resources led to an assembly with very few gaps. In contrast, other species were sequenced using the whole-genome shotgun approach, which produces fragmented assemblies that are far from chromosome-level. In addition, the only non-*elegans Caenorhabditis* species with significant genetic markers tied to the physical map is

C. briggsae [18–20]. The fragmentation of genomes immediately reduces the ability to use local and global synteny information to inform homology determination.

2.2 The Interaction of Mating System, Heterozygosity, and Gene Number

C. elegans, *C. briggsae*, and *C. tropicalis* (formerly *C.* sp. 11; [21]) produce self-fertile hermaphrodites. As a result, individuals are often completely homozygous in nature, and no inbreeding depression is seen in these species [22–24]. Being naturally devoid of allelic variation, even nearly identical sequences can confidently be recognized as duplications, even if they are on different sequence contigs. This is not true, however, of obligately outcrossing *Caenorhabditis*, such as *C. remanei* [25]. These organisms have high levels of allelic divergence, up to 14 % at silent positions in *C. brenneri* [26]. Moreover, among the allelic variants are abundant recessive deleterious mutations, which lead to severe inbreeding depression [22].

The presence of abundant recessive deleterious mutations, which is normally tolerated in the large populations of outcrossing *Caenorhabditis*, creates a special challenge for genome sequencing. Inbreeding is the simplest way to reduce heterozygosity in a genome, but is less effective than expected in outcrossing species due to balanced polymorphisms [27]. In some cases large fractions of the genome can remain heterozygous. In combination with automated gene annotation, many redundant allelic variants are added that appear to the end-user as highly similar paralogs. In addition, diverged haplotypes can contribute to fragmentation of other parts of the genome by partitioning sequence reads into two competing assemblies. *Post hoc* distinction of alleles from recently diverged paralogs is nontrivial, as the most abundant class of *bona fide* gene duplicates is the most recent [28], so that no simple sequence identity threshold suffices to parse duplicates from alleles. After the initial demonstration of the severity of heterozygosity in outcrossing *Caenorhabditis* genome assemblies [27], improved pipelines were developed [26, 29] to systematically recognize (and even exploit) such regions. However, as of this writing the depiction of unrecognized alternative alleles as paralogs remains an issue.

Mating system also impacts the size and gene content of genome assemblies in a surprisingly predictable fashion. The number of gene predictions for obligately outcrossing species is consistently larger than for selfing species, and fewer distinct mRNAs are detected in their transcriptomes [17]. This is likely due to the interaction between relaxed selection on mating-related traits [30, 31] and sex-biased transmission of chromosomes that favors loss of larger size variants by drift in partially selfing species [32].

2.3 Phylogeny vs. Raw Similarity

In an idealized world of molecular evolution in which gene loss and duplication do not occur, and rates of molecular evolution are even across a gene family, orthologs will show a greater similarity

than any other possible inter-species pairings. In these cases, simple heuristics for inferring orthologs, such as reciprocal best BLAST hits [14, 33], will generally identify the ortholog, even if there are related paralogs as well. However, it is common for rates of molecular evolution to be uneven, and in some gene families gain and loss are commonplace. In extreme cases, little orthology remains despite the presence of consistently large gene families (e.g. [34]). This is often only clear after inferring a phylogenetic tree. In WormBase, most *Caenorhabditis* homologs of *C. elegans* genes have been run through automated pipelines (reviewed in ref. 35), such as TreeFam [36], Imparanoid [33], and OrthoMCL [37]. Their outputs, in combination with more focused analyses, are currently presented in the "curated nematode orthologs" section of the WormBase gene page. These tools represent major advances over raw similarity, but should nevertheless be regarded as preliminary. Particularly difficult to recognize are cases of parallel and/or complementary loss of members of established subfamilies (Fig. 1). For example, the most similar *C. briggsae* genes to *C. elegans fbf-1* and *fbf-2* are *Cbr-puf-1*, *Cbr-puf-1.2*, *and Cbr-puf-2*, and they were initially inferred to be orthologous [38]. However, the inclusion of homologs from *C. japonica*, *C. remanei*, and *C. brenneri* revealed that they were actually members of two distinct and ancient PUF subfamilies that had experienced reciprocal losses in the *C. elegans* and *C. briggsae* lineages [39]. For such cases, careful alignment and appropriate taxon sampling is essential to clarify the situation (*see* Subheading 4 below).

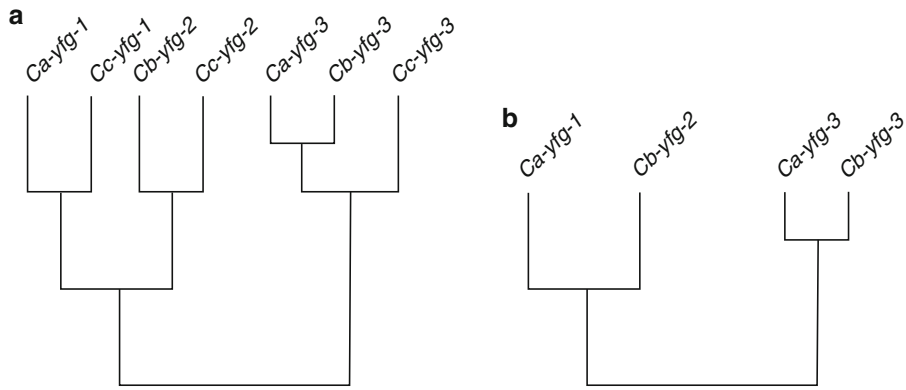

Fig. 1 Misleading effects of insufficient taxon sampling. We consider three hypothetical *Caenorhabditis* species, *Ca, Cb, and Cc*. (**a**) The last common ancestor of these three species had three paralogs of *y*our *f*avorite *g*ene, *yfg-1, yfg-3,* and *yfg-3*. However, species *Cb* has lost its *yfg-1* paralog, and *Ca* its *yfg-2*. (**b**) When species *Cc* is omitted from consideration, however, the complementary losses are not apparent, and *Ca-yfg-1* and *Cb-yfg-2* could be erroneously inferred to be orthologs

3 A Decision Tree for Assessing Gene Homology in Nematode Genome Assemblies

Definitive studies of homology relationships are time-consuming, and still form entire research publications. However, an experimental biologist not trained in molecular phylogenetics often needs a "quick and dirty" way to sift through various homologs that is still defensible. In this spirit, we present a sort of key, similar in layout to that used in species identification, to help guide scrutiny of homologs. The assumption is that the researcher will be starting with a single *C. elegans* query gene, which may or may not be part of a family, and that the goal is to identify the orthologous gene (if one exists) for a specific second species among the potentially numerous homologs listed in WormBase gene pages. It also provides a way to quickly verify that an annotated ortholog really is orthologous.

1) *Zero homologs.* The complete absence of related sequences in gene predictions for another *Caenorhabditis* assembly is rare, but can happen for a number of reasons:
   (a) A homologous sequence may be present in the assembly, but is not annotated. This possibility can be tested by searching the genomic DNA with the protein sequence of the *C. elegans* using TBLASTN [40], which compares the protein query to conceptual translations, in all six possible reading frames, of the genome.
   (b) There may be a homolog, but it falls in a gap in sequence coverage. In this case, examination of syntenic genes may reveal the gap.
   (c) The absence may be a biological reality. As noted above, thousands of genes were likely lost in the *C. elegans* genome after the transition to self-fertility. Genes specific to *C. elegans* are rarer, but exist and can even encode essential factors [41].
2) *Exactly one homolog.* This single homolog is likely to be the ortholog, barring unrecognized homologs in sequence or annotation gaps. Conserved flanking genes (local synteny; [42]) often (but not always) confirm the diagnosis.
3) *Exactly two homologs.*
   (a) The two homologs are nearly identical to each other (e.g. over 90 % at the nucleotide level). These may be recent duplicates or retained alleles.
      (i) If allelic:

(1) The two homologs will be on different sequence contigs, one of which is usually short (<15 kb).

(2) Flanking genes that are single-copy in *C. elegans* will tend to also be present in two copies in this species.

(3) The two homologs are autosomal (male hemizygosity of the X prevents recessive deleterious alleles from accumulating).

(4) Noncoding sequences (introns, 3′ UTR, flanking DNA) will align easily, though with some insertions and deletions (e.g. [43]).

(ii) If recently duplicated paralogs:

(1) The two homologs are likely to be in tandem or near each other on the same large sequence contig.

(2) The two homologs may be X-linked.

(3) Noncoding sequences may be highly diverged.

(4) The two homologs are more likely to be in the high-recombination peripheral domains than in the central "cluster" domain [20, 44].

(b) The two homologs are quite divergent, e.g. over 10 % at the amino acid level. They are likely to be duplicates (paralogs).

(i) The *C. elegans* query also has two or more diverged copies. The duplication may have occurred before the last common ancestor of it and the other species. A simple distance-based phylogeny (*see* 4 Phylogeny Tips below) could confirm or reject this.

(ii) The *C. elegans* query has no within-species paralog. This would suggest that either *C. elegans* lost a copy, or that the other species experienced a duplication after it diverged from their common ancestor. Adding homologs from a more distantly related species and creating a sequence-based phylogeny usually clarifies which is the case.

(4) *Three or more homologs.* Allelism is an insufficient explanation; gene copy number has definitely evolved.

(1) The *C. elegans* query has two or more related sequences in *C. elegans* (*see* 2bi above). Many cases of large multigene families exist in nematode genomes (e.g. [45]), and most require a phylogenetic approach to understand.

(2) The *C. elegans* query has no within-species paralog (i.e. a 1:3+ situation). Either expansion or contraction has occurred in one of the species (*see* 2bii above).

4 Phylogeny Tips

As noted above, when a many-to-many relationship exists between related sequences in different species, a phylogeny (or "gene tree") is the only way to achieve clarity. A full treatment of how to infer molecular phylogenies is beyond the scope of this paper, and we refer the reader to other sources (e.g. [46, 47]). However, we can offer here some simple rules of thumb for nematode sequences that will greatly increase the utility of trees obtained:

- *Use protein sequences (not coding DNA) to infer the tree.* Between most *Caenorhabditis* species there has been complete mutational saturation of silent sites [48], whose inclusion would thus only introduce noise into the analysis.
- *Only confidently alignable sequences should be used for the analysis.* Identity need not be consistently high if there are conserved motifs to serve as landmarks (e.g. [49]).
- *Inclusion of more taxa often produces a better result.* For example, though the most similar *C. briggsae* genes to *C. elegans fbf-1* and *fbf-2* are *Cbr-puf-1*, *Cbr-puf-1.2*, *and Cbr-puf-2*, and were initially inferred to be orthologous [38] the inclusion of homologs from *C. japonica*, *C. remanei*, and *C. brenneri* revealed that they were actually members of two distinct and ancient PUF subfamilies that had experienced reciprocal losses in the *C. elegans* and *C. briggsae* lineages [39].

5 Scaling Up

Most of the tests described above can be performed for a handful of genes with little trouble using only the WormBase web site, and there are many suitable software tools that can be used to infer phylogenies from sequence alignments. However, in some cases finding homologs, if not necessarily orthologs, of large sets of query genes (e.g. >100) is desirable. In these cases, we offer the following compendium of specific methods.

5.1 Genome-Scale Sequence Datasets

Working with large sets of genes, or genome wide is possible thanks to resources available on the ftp website of WormBase (ftp://ftp.wormbase.org/pub/wormbase/; also linked at the bottom of the WormBase main page). For each release of WormBase, text files containing the data upon which the website is built are available for

download. This is particularly useful for species other than *C. elegans*, for which data are not necessarily retrievable through the use of WormMart or WormMine or do not exist yet (e.g. BLASTP hits or orthology assignments), or when analyzing a large set of genes. In addition, for reasons stated above, it might sometimes be preferable or necessary to determine orthology relationships for a family of related sequences independently.

The number of sequence files for each species varies depending on the status of the genomic assembly and annotations, availability of expression data, and curation of any given species at each WormBase release. All genome draft assemblies, except for those of *C. elegans* and *C. briggsae*, are fragmented in numerous scaffolds. The filenames are generally self-explanatory, and for the purpose of identifying homologs only a few are useful: genome draft sequences are stored in FASTA-format files (plain text with a specific header line for each sequence) ending with "genomic.fa," and if they exist, coding sequences in those ending with "cds_transcripts.fa." Annotations of the genomic scaffolds, including the positions of exons and introns of gene predictions, end with "annotations.gff3" or "annotations.gff2."

5.2 Running Batch BLAST Searches from the Command Line

Identifying allelic scaffolds (*see* above) can be a first step in sorting potential gene paralogs from allelic variants. Once these have been recognized, orthologs are often provisionally identified as best reciprocal BLAST hits. When dealing with a large set of genes, this process can be automated with command-line BLAST [50], which can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. Once installed onto a desktop or laptop computer, command-line searches are initiated through the Terminal (Unix or Mac) or Command Prompt (Windows) applications. This is most easily done by placing the relevant BLAST databases and query files in the same, or in closely linked, directories.

Though less familiar to many experimental biologists than web-based BLAST servers, there are major advantages to using command-line BLAST. One is flexibility. Any number of CDS sequences can be input as queries in a batch, up to the entire set of predictions for a given species. This set of sequences can be compared to either another set of gene predictions, filtered or not to remove allelic variants, or to genomic sequences. The latter case should generate more than one high-confidence hit per CDS sequence, but in simple cases only one per exon. Another advantage to command-line BLAST is the ability to quickly vary the values of the parameters for BLAST as well as for subsequent filtering. The threshold above which a hit is considered significant might vary depending on the accuracy and completeness of the assembly or set of CDS sequences, as well as on the degree of divergence between species and level of heterozygosity found within each species. The latter especially matters when using sequences from outcrossing species generated

in-house (rather than those stored on WormBase), since the degree and location of heterozygous regions is unknown, and more than one individual contributes their DNA to sequencing.

As noted above, a sensible BLAST *E*-value threshold between best reciprocal hits is not enough to indicate orthology. When working with large gene sets, additional parameters to be considered include the proportion of each gene's length involved in the alignment, the degree of identity over the length of the alignment, and for close relatives, conservation of micro-synteny. Additionally, in the case of genomic sequences hits, the alignments corresponding to potential exons must be colinear to the query sequence. Once the output files have been generated, BLAST alignments can be sorted through efficiently with appropriate Perl or Python scripts.

5.3 Identifying All Members of a Gene Family

Another common task is to identify the entire set of genes in a family, as defined by possession of one or more protein domains, a specific motif, or a highly conserved region. In these cases, a number of options are available. For short motifs that can be represented by a simple regular text expression, Perl or Python scripts can be used to search sequences directly. In more complex cases, a tool like HMMER [51, 52] can generate a profile from the shared features, and be used to comb through CDS or genomic sequence databases to identify occurrences. The same considerations as for BLAST searches apply here as well to sensibly determine homology. Phylogenetic analysis of the genes' relationships to one another is required to assign accurately orthology.

References

1. The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018
2. Colaiacovo MP, Stanfield GM, Reddy KC, Reinke V, Kim SK, Villeneuve AM (2002) A targeted RNAi screen for genes involved in chromosome morphogenesis and nuclear organization in the *Caenorhabditis elegans* germline. Genetics 162(1):113–128
3. Gonczy P, Echeverri C, Oegema K, Coulson A, Jones SJ, Copley RR, Duperon J, Oegema J, Brehm M, Cassin E, Hannak E, Kirkham M, Pichler S, Flohrs K, Goessen A, Leidel S, Alleaume AM, Martin C, Ozlu N, Bork P, Hyman AA (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. Nature 408(6810): 331–336
4. Simmer F, Moorman C, van der Linden AM, Kuijk E, van den Berghe PV, Kamath RS, Fraser AG, Ahringer J, Plasterk RH (2003) Genome-wide RNAi of *C. elegans* using the hypersensitive *rrf-3* strain reveals novel gene functions. PLoS Biol 1(1), E12
5. Edgley M, D'Souza A, Moulder G, McKay S, Shen B, Gilchrist E, Moerman D, Barstead R (2002) Improved detection of small deletions in complex pools of DNA. Nucleic Acids Res 30(12), e52
6. Boulin T, Etchberger J, Hobert O (2006) Reporter gene fusions. In: WormBook, The *C. elegans* Research_Community, Editor 2006
7. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dose AC, Du

J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM III, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Ratsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330(6012): 1775–1787

8. Reinke V, Gil IS, Ward S, Kazmer K (2004) Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. Development 131(2):311–323
9. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS (2001) A gene expression map for *Caenorhabditis elegans*. Science 293(5537): 2087–2092
10. Reinke V, Smith HE, Nance J, Wang J, Van Doren C, Begley R, Jones SJ, Davis EB, Scherer S, Ward S, Kim SK (2000) A global profile of germline gene expression in *C. elegans*. Mol Cell 6(3):605–616
11. Haag E, Pilgrim D (2005) Harnessing Caenorhabditis genomics for evolutionary developmental biology. Curr Genomics 6:579–588
12. Butler MH, Wall SM, Luehrsen KR, Fox GE, Hecht RM (1981) Molecular relationships between closely related strains and species of nematodes. J Mol Evol 18(1):18–23
13. Zucker-Aprison E, Blumenthal T (1989) Potential regulatory elements of nematode vitellogenin genes revealed by interspecies sequence comparison. J Mol Evol 28(6): 487–496
14. Stein L et al (2003) The genome sequence of *Caenorhabditis briggsae:* a platform for comparative genomics. PLoS Biol 1:166–192
15. Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, Kato M, Miller DM, Slack F, Snyder M, Waterston RH, Reinke V, Gerstein MB (2011) Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res 21(2):276–285
16. Jovelin R, Cutter AD (2013) Fine-scale signatures of molecular evolution reconcile models of indel-associated mutation. Genome Biol Evol 5(5):978–986
17. Thomas CG, Li R, Smith HE, Woodruff GC, Oliver B, Haag ES (2012) Simplification and desexualization of gene expression in self-fertile nematodes. Curr Biol 22:2167–2172
18. Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, Waterston RH (2007) Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol 5(7), e167
19. Koboldt DC, Staisch J, Thillainathan B, Haines K, Baird SE, Chamberlin HM, Haag ES, Miller RD, Gupta BP (2010) A toolkit for rapid gene mapping in the nematode *Caenorhabditis briggsae*. BMC Genomics 11:236
20. Ross J, Koboldt D, Staisch J, Chamberlin H, Gupta BP, Millller R, Baird S, Haag E (2011) *Caenorhabditis briggsae* recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination. PLoS Genet 7(7), e1002174
21. Félix M-A, Braendle C, Cutter AD (2014) A streamlined system for species diagnosis in Caenorhabditis (Nematoda: Rhabditidae) with name designations for 15 distinct biological species. PLoS One 9(4), e94723
22. Dolgin ES, Charlesworth B, Baird SE, Cutter AD (2007) Inbreeding and outbreeding depression in Caenorhabditis nematodes. Evolution 61(6):1339–1352
23. Barriére A, Félix MA (2005) High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. Curr Biol 15(13):1176–1184
24. Cutter AD, Felix MA, Barriere A, Charlesworth D (2006) Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of *Caenorhabditis briggsae*. Genetics 173(4): 2021–2031
25. Cutter AD, Baird SE, Charlesworth D (2006) High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. Genetics 174(2):901–913

26. Dey A, Chan CK, Thomas CG, Cutter AD (2013) Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. Proc Natl Acad Sci U S A 110(27): 11056–11060
27. Barriere A, Wang S, Pekarek E, Thomas C, Haag E, Ruvinsky I (2009) Detecting heterozygosity in shotgun genome assemblies: lessons from obligately outcrossing nematodes. Genome Res 19:470–480
28. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155
29. Huang S, Chen Z, Huang G, Yu T, Yang P, Li J, Fu Y, Yuan S, Chen S, Xu A (2012) HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res 22(8):1581–1588
30. Cutter AD (2008) Reproductive evolution: symptom of a selfing syndrome. Curr Biol 18(22):R1056–R1058
31. Thomas CG, Woodruff GC, Haag ES (2012) Causes and consequences of the evolution of reproductive mode in Caenorhabditis nematodes. Trends Genet 28(5):213–220
32. Wang J, Chen PJ, Wang GJ, Keller L (2010) Chromosome size differences may affect meiosis and genome size. Science 329(5989):293
33. Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer EL (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38(Database issue):D196–D203
34. Nayak S, Goree J, Schedl T (2005) *fog-2* and the evolution of self-fertile hermaphroditism in Caenorhabditis. PLoS Biol 3, e6
35. Altenhoff A, Dessimoz C (2012) Chpt. 9: Inferring orthology and paralogy. In: Anisimova M (ed) Evolutionary genomics, statistical and computational methods, vol 1. Humana Press, New York, pp 259–279
36. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34(Database issue): D572–D580
37. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9): 2178–2189
38. Lamont LB, Crittenden SL, Bernstein D, Wickens M, Kimble J (2004) FBF-1 and FBF-2 regulate the size of the mitotic region in the *C. elegans* germline. Dev Cell 7(5):697–707
39. Liu Q, Stumpf C, Wickens M, Haag ES (2012) Context-dependent function of a conserved translational regulatory module. Development 139:1509–1521
40. Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
41. Zhang T, Sun Y, Tian E, Deng H, Zhang Y, Luo X, Cai Q, Wang H, Chai J, Zhang H (2006) RNA-binding proteins SOP-2 and SOR-1 form a novel PcG-like complex in *C. elegans*. Development 133(6):1023–1033
42. Kuwabara PE, Shah S (1994) Cloning by synteny: identifying *C. briggsae* homologues of *C. elegans* genes. Nucleic Acids Res 22: 4414–4418
43. Haag ES, Ackerman AD (2005) Intraspecific variation in *fem-3* and *tra-2*, two rapidly coevolving nematode sex-determining genes. Gene 349:35–42
44. Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLoS Genet 5(3), e1000419
45. Thomas JH, Kelley JL, Robertson HM, Ly K, Swanson WJ (2005) Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Proc Natl Acad Sci U S A 102(12):4476–4481
46. Swofford D, Olsen G, Waddell P, Hillis D (1996) Chpt. 11: Phylogenetic inference. In: Hillis D, Moritz C, Mable B (eds) Molecular systematics. Sinauer Associates, Sunderland, MA
47. Whelan S (2008) Inferring trees. In: Keith J (ed) Bioinformatics, vol 1, Data, sequence analysis, and evolution. Humana Press, Totowa, NJ, pp 287–309
48. Cutter AD, Dey A, Murray RL (2009) Evolution of the *Caenorhabditis elegans* genome. Mol Biol Evol 26(6):1199–1234
49. Haag ES, Wang S, Kimble J (2002) Rapid coevolution of the nematode sex-determining genes *fem-3* and *tra-2*. Curr Biol 12(23): 2035–2041
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
51. Eddy S (2001) HMMER: profile hidden Markov models for biological sequence analysis., Available from: http://hmmer.wustl.edu
52. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39(Web Server issue):W29–W37

http://www.springer.com/978-1-4939-2841-5

C. elegans

Methods and Applications

Biron, D.; Haspel, G. (Eds.)
2015, XII, 252 p. 57 illus., 44 illus. in color., Hardcover
ISBN: 978-1-4939-2841-5
A product of Humana Press