Chapter 5:

Incorporating Genomics Into the Toolkit of Nematology*

*This chapter first published in *Journal of Nematology* in 2012 and was written by Adler R. Dillman, Ali Mortazavi, and Paul W. Sternberg. This chapter includes discussion of *Steinernema* genomes for which a separate manuscript is being prepared but was not yet ready to include in this thesis.

Abstract

The study of nematode genomes over the last three decades has relied heavily on the model organism *Caenorhabditis elegans*, which remains the best-assembled and annotated metazoan genome. This is now changing as a rapidly expanding number of nematodes of medical and economic importance have been sequenced in recent years. The advent of sequencing technologies to achieve the equivalent of the $1000 human genome promises that every nematode genome of interest will eventually be sequenced at a reasonable cost. As the sequencing of species spanning the nematode phylum becomes a routine part of characterizing nematodes, the comparative approach and the increasing use of ecological context will help us to further understand the evolution and functional specializations of any given species by comparing its genome to that of other closely and more distantly related nematodes. We review the current state of nematode genomics and discuss some of the highlights that these genomes have revealed and the trend and benefits of ecological genomics, emphasizing the potential for new genomes and the exciting opportunities this provides for nematological studies.

Introduction

Nematoda is one of the most expansive phyla documented with free-living and parasitic species found in nearly every ecological niche [1]. Traditionally, nematode phylogeny was based on classical and often incomplete understanding of morphological traits, but traditional systems have been revised and supplemented by a growing body of insight from molecular phylogenetics that is primarily based on ribosomal DNA for higher level taxonomic studies [2–4]. The study of the evolutionary relationships between

species in vertebrates and in arthropods is transitioning to the comparative analysis of entire genomes due to the exponentially decreasing cost of sequencing and the study of nematodes is now following the same path [5–7]. While the model organism *Caenorhabditis elegans* was the first metazoan sequenced [8], there have been only a few additional nematodes sequenced until recently and many representative clades and ecological niches remain unexplored. There are several advantages to whole genome sequencing for nematology. The simplest and most obvious is that the complete genome harbors the full repertoire of genes that are the inherited common core of any given species. Furthermore, the genome contains the structural and regulatory elements that lie in and between genes, even if we cannot yet identify them all. The genome also provides the foundation for future experimentation such as transformation and RNA interference (RNAi). The genome is the natural framework for indexing and organizing the massive genetic content of species within a phylum. The genetic 'blueprint' represented by a genome may prove to be the most valuable and enduring piece of knowledge we can currently obtain for any particular life form [8].

As in many other fields of biology, the nematode *C. elegans* has proven invaluable as a model for genomic analysis, and thousands of investigators have contributed to our understanding of its 20,431 protein-coding genes [8, 9]. This is likely for the same reasons that make this hermaphrodite so powerful and useful in genetics: 1) its ease of culture, 2) its simple, rapid, invariant development, 3) many biological principles are universal, even if specific details are not, and 4) the more detailed our understanding of any biological phenomenon, the more interesting it tends to become [10]. While sequencing efforts have expanded exponentially as technology improves and

the cost continues to diminish, the finished *C. elegans* genome remains unrivaled in completeness compared to other metazoans. This is not likely to change, due partly to differences in technology but primarily because closing the remaining gaps in genomic sequence is a prolonged and expensive process with diminishing biological return [8]. The top-down approach of completing genome sequences by breaking the genome down into large, known fragments, which provide a physical map, and the subsequent sequencing of those fragments in their entirety, will probably not be common until new technologies sharply reduce the costs of finishing genomes.

Over the last two decades, sequencing technology has advanced from relying on the hierarchical sequencing and assembly of cloned fragments of DNA (i.e., automated Sanger sequencing as used in the *C. elegans* project), to the shotgun, high-throughput ~ 500 bp reads produced by 454 Roche sequencing and the even cheaper ≤ 150 bp reads produced by Illumina sequencing [11–13]. Due to the rapid pace of sequencing technology development and turnover, we will refer to the newer technologies as 'next-generation' (next-gen) technologies throughout rather than focus on any specific platform. These next-gen technologies are driven with the eventual goal to achieve a ≤ $1000 human genome to enable health applications. Given that the typical nematode genome is less than 1/15 of the size of the 3.2 Gb human genome (see Table 5.1 for nematode genome sizes), sequencing nematode genomes is already affordable and, as technology improves, could become monetarily negligible. Current next-gen technologies use DNA fragments of various size to generate sequence, which range from less than 500 bp up to ≤ 20 kb, and can produce either single or paired end reads (either one or both ends of prepared fragments can be sequenced (Figure 5.1)). Next-gen sequencing technologies

generate many more sequencing reads that have a higher error rate than traditional Sanger sequencing, but this is balanced by higher overall coverage (whereas 2 Gb of generated sequence would provide 20-fold coverage of a 100 Mb genome, 10 Gb of generated sequence would provide 100-fold coverage). When considering these sequencing technologies it is important to distinguish fragment size and read length as distinct variables that will affect the resulting assembly, because it is easy to sometimes conflate or combine these separate aspects. Fragment size refers to the length of the DNA insert, from which sequence will be generated either from one or both sides, while read length refers to how many base pairs are actually being sequenced from one or both sides of the fragment (Figure 5.1).

The hierarchical application of Sanger sequencing to the assembly of the *C. elegans* genome helped to facilitate completeness and circumvented the potential problems of long repeats, homopolymeric regions, and low G+C content, along with the community effort of researchers, which was crucial and is ongoing [8, 14]. Next-gen technologies are more affordable and allow for much higher fold coverage of genomes, leading to hundreds of millions of genomic reads. In contrast to the hierarchical approach previously used, the shotgun strategies in favor today are based on breaking the entire genome into many more small fragments. These require more computational effort to assemble into multigenic sized contigs, let alone chromosomes (Figure 5.1) [11]. The contiguity of the resulting draft genomes can be dramatically improved by library construction with inserts of larger but approximately known sizes as well as ‘jumping libraries’ (Figure 5.1C) [15, 16].

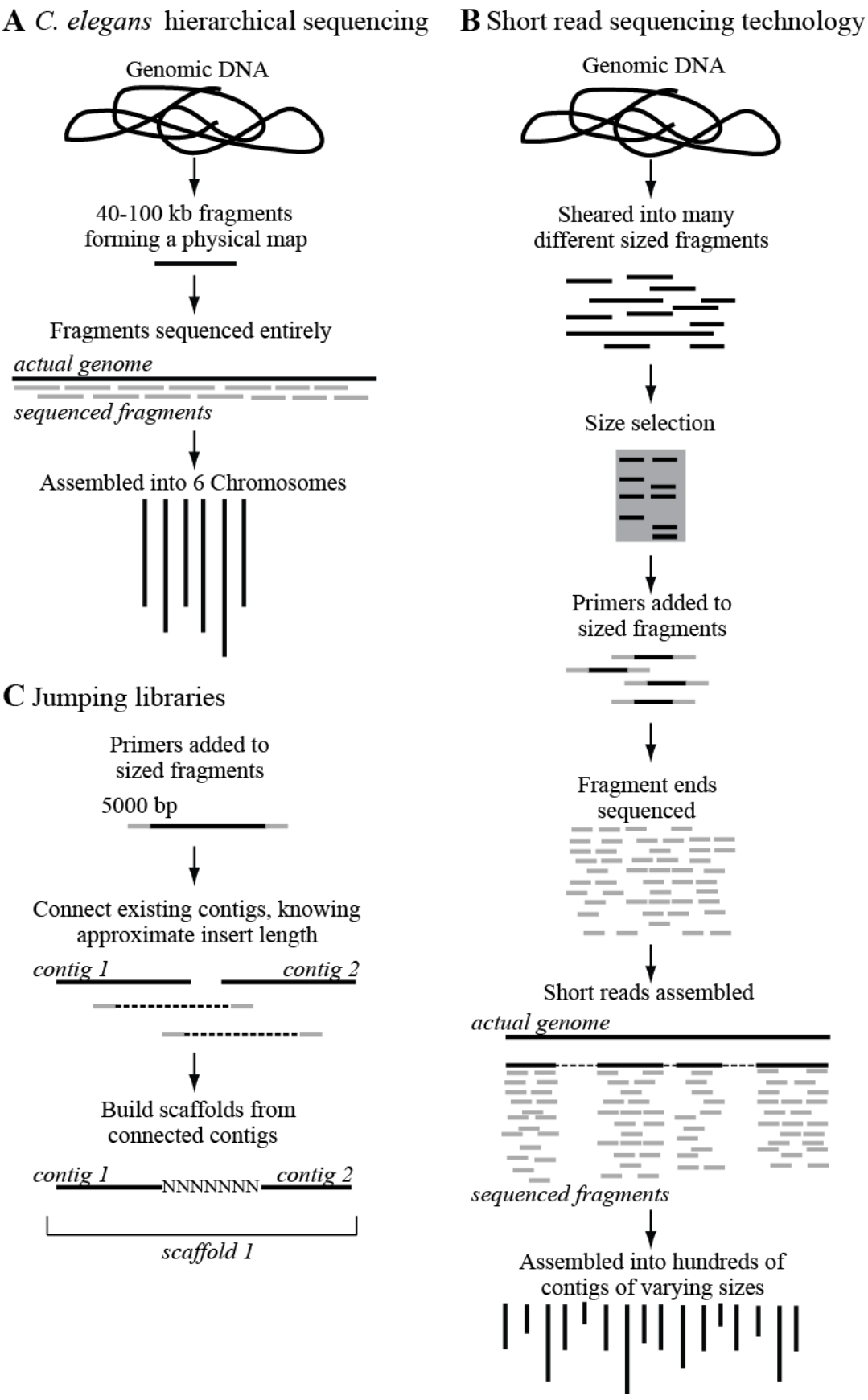
A C. elegans hierarchical sequencing
Genomic DNA
40-100 kb fragments
forming a physical map
Fragments sequenced entirely
actual genome
sequenced fragments
Assembled into 6 Chromosomes
B Short read sequencing technology
Genomic DNA
Sheared into many
different sized fragments
Size selection
Primers added to
sized fragments
Fragment ends
sequenced
Short reads assembled
actual genome
sequenced fragments
Assembled into hundreds of
contigs of varying sizes
C Jumping libraries
Primers added to
sized fragments
5000 bp
Connect existing contigs, knowing
approximate insert length
contig 1
contig 2
Build scaffolds from
connected contigs
contig 1
NNNNNNN
contig 2
scaffold 1

Figure 5.1 | Hierarchical and shotgun sequencing. A) A shortened diagram of the hierarchical (top-down) sequencing technique used for the *C. elegans* genome. The genomic DNA was broken into large fragments (40–100 kb) that formed a physical map. The order of the fragments was known before they were actually sequenced. The fragments were then fully sequenced and assembled, resulting in six chromosomal contigs. B) A diagram of the shotgun-sequencing techniques used to prepare genomic DNA for sequencing using 454 Roche or Illumina short-read technology. The genomic DNA is sheared into approximately sized fragments of 0.5 to 1 kb. These fragments then have primers attached to one or both ends, depending on whether they will be run on a paired-end sequencer. The fragments are not sequenced in their entirety, but 50–500 bp of one or both ends of each fragment are sequenced. The resulting short reads are then assembled, with some gaps remaining as shown (gaps are represented by the dotted lines). These reads are then assembled into hundreds up to thousands of larger contigs (contiguous sequence). C) Jumping libraries are used in short-read sequencing to improve assembly quality. During the size selection, larger fragment sizes are selected and sequenced. Only one size is selected per library, but the sizes range from 2 kb to 20 kb. Read assembly is facilitated by knowing the approximate distance between the paired end reads, helping to overcome issues of repeats and homopolymeric regions, jumping large regions (as large as the insert length). Assembly quality is improved as multiple previously unconnected contigs are now known to connect, just as contig 1 and contig 2 are joined to form a larger contig called ‘scaffold 1’ in the figure.

Since the first nematode genome was first published in 1998, twelve more whole nematode genomes have been sequenced and made publicly available [8, 12, 17–25]. There are at least 13 more nematode genomes scheduled for release in 2012, and several others in preparation (Table 5.1) [26]. Because Nematoda is so ecologically diverse and

species-rich (1 to 10 million species [27, 28]), phylogenetic relationships along with human health and agricultural considerations should inform sequencing efforts.

Nematode Species	Date Published	Size (Mb)	# Protein-coding genes	G+C Content
Ascaris suum	Oct. 2011	273	18,542	38%
Brugia malayi	Sept. 2007	96	21,252	31%
Bursaphelenchus xylophilus	Sept. 2011	75	18,074	40%
Caenorhabditis angaria	Oct. 2010	80	22,851	36%
*Caenorhabditis brenneri**	Feb. 2009	190	30,670	39%
Caenorhabditis briggsae	Nov. 2003	106	21,963	37%
Caenorhabditis elegans	Dec. 1998	100	20,431	35%
*Caenorhabditis japonica**	Feb. 2009	83	25,879	40%
*Caenorhabditis remanei**	Feb. 2009	145	31,471	38%
Caenorhabditis sp. 5†	~2012	132	35,000	39%
Caenorhabditis sp. 11†	~2012	79	22,330	38%
Dictyocaulus viviparus§	-	230	-	-
Dirofilaria immitis§	-	90	-	-
Haemonchus contortus†	~2012	315	-	42%
Heterorhabditis bacteriophora†	~2012	74	-	33%
Heterorhabditis indica†	~2012	64	-	34%
Heterorhabditis megidis†	~2012	71	-	34%
Heterorhabditis sonorensis†	~2012	64	-	34%
Litmosoides sigmodontis§	-	90	-	35%
Loa loa§	-	90	-	-
Meloidogyne hapla	Sept. 2008	53	13,072	27%
Meloidogyne incognita	Jul. 2008	86	19,212	31%
Nippostrongylus brasiliensis§	-	80	-	-
Onchocera ochengi§	-	90	-	-
Onchocera volvulus§	-	90	-	-
Oscheius tipulae§	-	120	-	-
Panagrellus redivivus†	~2012	62	27,266	44%
Pristionchus pacificus	Sept. 2008	173	24,216	43%
Steinernema carpocapsae†	~2012	86	-	46%
Steinernema feltiae†	~2012	101	-	47%
Steinernema glaseri†	~2012	94	-	48%
Steinernema monticolum†	~2012	114	-	45%
Steinernema scapterisci†	~2012	86	-	48%
Strongyloides ratti§	-	50	-	-
Trichinella spiralis	Mar. 2011	64	15,808	34%
Wuchereria bancrofti§	-	90	-	-

Table 5.1 | Genome statistics for published and selected forthcoming whole nematode genomes. Additional genomes planned and in progress can be viewed at www.nematodes.org/nematodegenomes/index.php/959_Nematode_Genomes.

* Genomes that were not explicitly published in genome papers but discussed in Barrière *et al*. [22].

† Forthcoming genome statistics were provided by investigators working on those projects (P.W. Sternberg, E.M. Schwarz, and H.T. Schwartz).

§ Genomes in production with data available from the 959 genome project website.

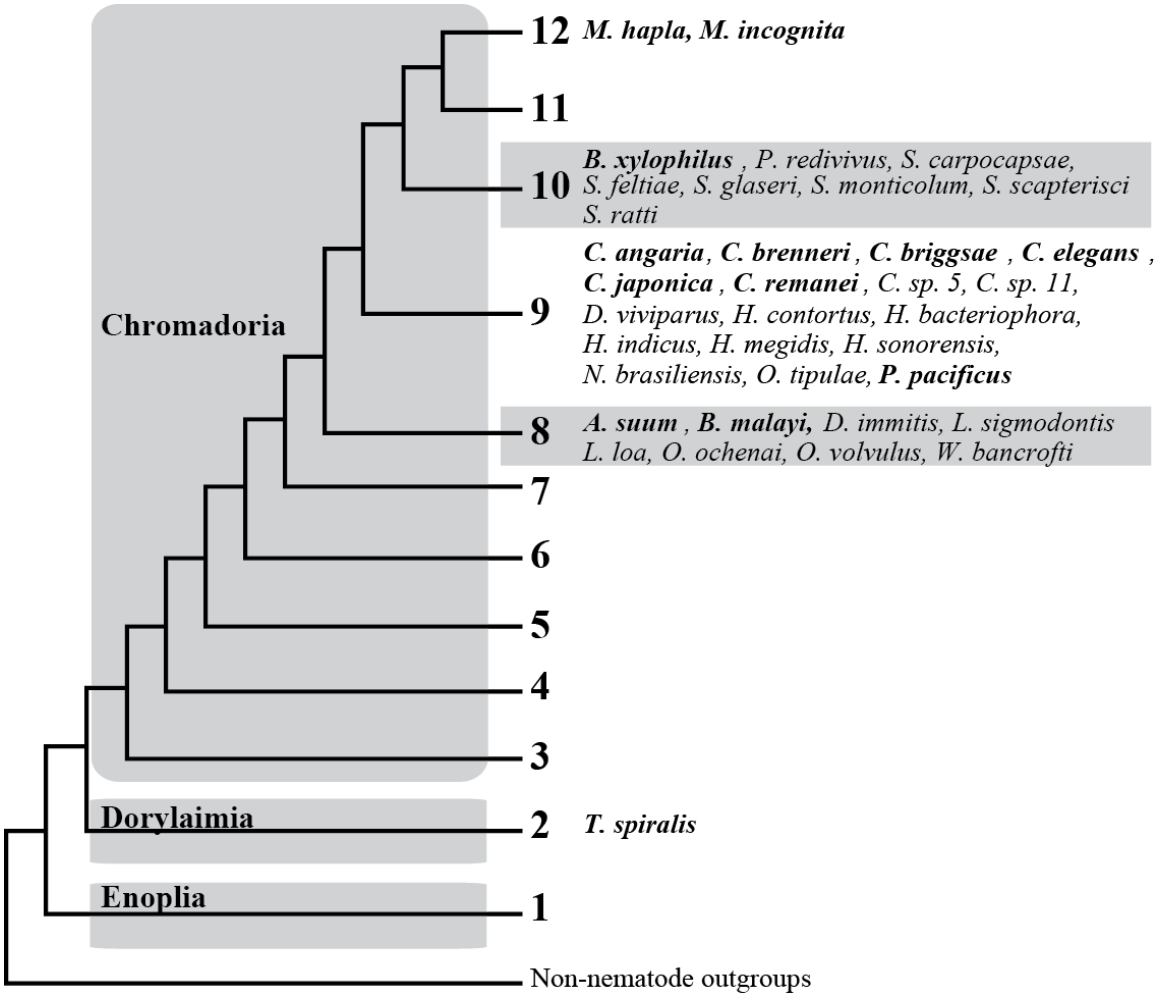

Figure 5.2 | A phylogenetic representation of sequenced and selected forthcoming nematode genomes. This diagram depicts twelve monophyletic clades representing the phylum Nematoda. Sequencing efforts have focused on a few select crown clades of Chromadoria. The clade designations are after Holterman *et al.* [4]. Taxa with published genomes have bolded names while taxa for which genomes are underway and scheduled for release in 2012 are regular typeface. Full genus names can be found in Table 1. Forthcoming genomes were selected from genomes with data available from the 959 genome project [26] and from genome projects of which the authors had knowledge.

The current view of nematode genealogical relationships divides the phylum into 3 major clades: Enoplia, Dorylaimia, and Chromadoria [2, 3, 29]. Chromadoria is further broken down into 10 clades, which together with Enoplia and Dorylaimia form a total of 12 major monophyletic branches within the phylum (Figure 5.2) [4, 6, 30]. Sequencing efforts so far have focused on nematodes in the crown clades of Chromadoria, which include *C. elegans* as well as most medically and agriculturally relevant species (Figure 5.2). A systematic genomic survey of the phylum would facilitate a better understanding of the evolution of Nematoda, enhance comparative studies, and could illuminate striking differences across the phylum such as differences in parasitic lifestyle (e.g., endoparasitic vs. ectoparasitic) or mode of reproduction (e.g., amphimictic vs. parthenogenetic) as well as developmental differences (e.g., asymmetric vs. symmetric cleavages; presence vs. absence of a prominent coeloblastula [31]), among others.

What are the benefits of genomics for nematologists? Herein we briefly review the basic information provided by most nematode genome analyses. We discuss the highlights of the 13 available nematode genomes, how their utility increases as the number of possible comparisons increases, and how the focus of nematode genomics is changing to emphasize the specific biology and ecology of each species. We finish by illustrating the potential benefit of sequencing additional nematode genomes, using as an example the prospects of entomopathogenic nematode genomes and discussing how they can contribute to our understanding of parasitism, mutualism, and nematode biology in general.

The steps in sequencing a genome

With such a diversity of nematodes to choose from, which nematodes should be sequenced first? In addition to the above-mentioned biological motivations of phylogenetic position and human health and agricultural concerns, there are practical considerations such as the availability and homogeneity of material. Culturability is also a consideration, especially if investigators are interested in the transcriptome and subsequent experimentation. Adding transcriptional data can dramatically improve gene predictions and assembly quality [23, 32]. Whole-genome amplification techniques may make it possible to analyze interesting-but-unculturable nematodes in a cost-effective way. However, such amplification techniques may introduce additional problems such as polymorphisms and amplification errors, while culturable worms escape these difficulties since they can provide large quantities of DNA (typically 5 micrograms are needed to construct robustly a representative DNA library, which corresponds to ~ 50,000 worms for *C. elegans*) and can be inbred to decrease heterozygosity. While the study of sequence variation within a species is of great importance, the same variation can make it difficult to assemble a genome *de novo* without producing assembly errors. Therefore every effort should be undertaken, if possible, to inbreed the strains used, to minimize polymorphisms. The genomic value of a culturable worm increases with complementary transcriptome data and the possibility of further experimentation. In fact, the implementation of some experimental techniques such as RNAi may depend on optimized culturing techniques that do not stress the nematodes being cultured [33]. We believe that there are plenty of interesting culturable nematodes that can shed light on the evolution of the phylum and thus should be prioritized to fill sequencing pipelines. While the bulk of our discussion

below focuses on genomic libraries, RNA-seq libraries for transcriptome sequencing can be built from as little as 100 ng of total RNA thus lowering the numbers of worms needed to collect data. As next-gen technologies mature, we can expect that the starting amounts of material necessary will decrease.

Once a suitable nematode is identified, the simplified, general pipeline for genomic sequencing is as follows: 1) extraction and purification of genomic DNA, 2) selection of a sequencing platform, 3) library construction, 4) sequencing, 5) assembly of the sequence into as long and as few contigs as possible, 6) gene predictions and subsequent annotation.

1) DNA extraction and purification. There are numerous DNA extraction and purification methods and proprietary kits that have been tested and are known to work well both for populations and individual nematodes [34–36].

2) Selection of a sequencing platform. Careful consideration should be given to selecting the appropriate sequencing technology and accompanying parameters, such as read length and fragment size. A common priority is to select the most cost-effective source of high-quality sequence while simultaneously collecting as many reads as possible to ensure good coverage. Good assemblies with short-read technologies typically require 100x average coverage to compensate for high error rates. Coverage takes into account the size of the genome and the length of sequenced reads; for a 100 Mb genome, 100 million 100 bp reads are needed to achieve 100x coverage. Matters are further complicated by the effect of GC-content (GC content of the genome is the percentage of guanines and cytosines) on the coverage in some next-gen technologies, which necessitate greater overall sequencing depth (i.e., more sequencing reads) to cover GC-

poor regions well [23]. Certain sequencing platforms may be advisable for particularly GC-poor genomes (e.g. < 35%), such as 454.

3) Library construction. Good library construction is often a critical step, depending on the sequencing technology used [37]. A genomic library is essentially genomic DNA that has been sheared into fragments, which are then size selected for an approximate distribution. These fragments then have sequencing primers ligated to one or both ends (Figure 5.1). Because of the massive number of reads, and increasingly longer read lengths, the construction of good libraries with a normally distributed fragment size can make the difference between good and poor quality assemblies. Libraries with average fragment sizes of 500 bp are sufficient to assemble most nematode-size gene loci onto a single contig [32]. Genomes that are rich in longer repeat sequences or gene clusters that are larger than the fragment lengths will benefit from additional jumping libraries, which are paired-end libraries that are typically 3–20 kb apart (Figure 5.1C) [12]. In addition to traditional genomic jumping libraries, transcriptome data can be used to scaffold expressed genes that are broken across multiple contigs [23].

4) Sequencing. After a library is constructed, it is then sequenced, which is typically handled by dedicated facilities. The sequencing run may take 1 to 10 days, but this may be prolonged depending on facility scheduling considerations. The resulting raw reads each consist of a DNA sequence and a corresponding quality score; these can be used to filter all but the highest-quality reads, which will improve the overall assembly.

5) Genome assembly. Reads are assembled into contigs using one of several available programs such as Velvet and SOAPdenovo [38, 39]. Genome assembly is a resource-intensive step that can require substantial memory, but the relatively small size

of nematode genomes makes assembly practical on servers with 128 to 256 gigabytes of RAM. Assembly programs work by finding overlap between reads into contigs and by connecting contigs using the paired information from paired-end (or jumping libraries) into scaffolds (connected contigs). In an ideal situation, one contig or even one scaffold per chromosome would be recovered, but this has only been achieved for *C. elegans* and *C. briggsae* (Figure 5.1A) [8, 17]. Assembly programs are often run multiple times with different parameters to maximize several of the assembly metrics described in the basic genome statistics section below.

6) Gene prediction and genome annotation. Once reads have been assembled, gene-finding programs that identify protein coding or non-protein coding genes such as Augustus and tRNAscan are used to annotate the genome (Figure 5.3) [40, 41]. Perhaps the most helpful additional dataset for this step is transcriptome data that is generated by high-throughput sequencing of mRNA (RNA-seq). This provides expression data and identifies *bona fide* transcripts (either full length or fragments) directly. These data can also be used to train prediction software, thus facilitating more reliable gene predictions [23, 32]. The transcriptome provides interesting biological data about global gene expression and can be applied to nematodes at specific stages such as infective juveniles or embryos. RNA-seq data for any biological sample, whether strain (e.g., drug-resistant mutant compared to the wild type) or stage-specific, can be used to identify genes with expression patterns of interest.

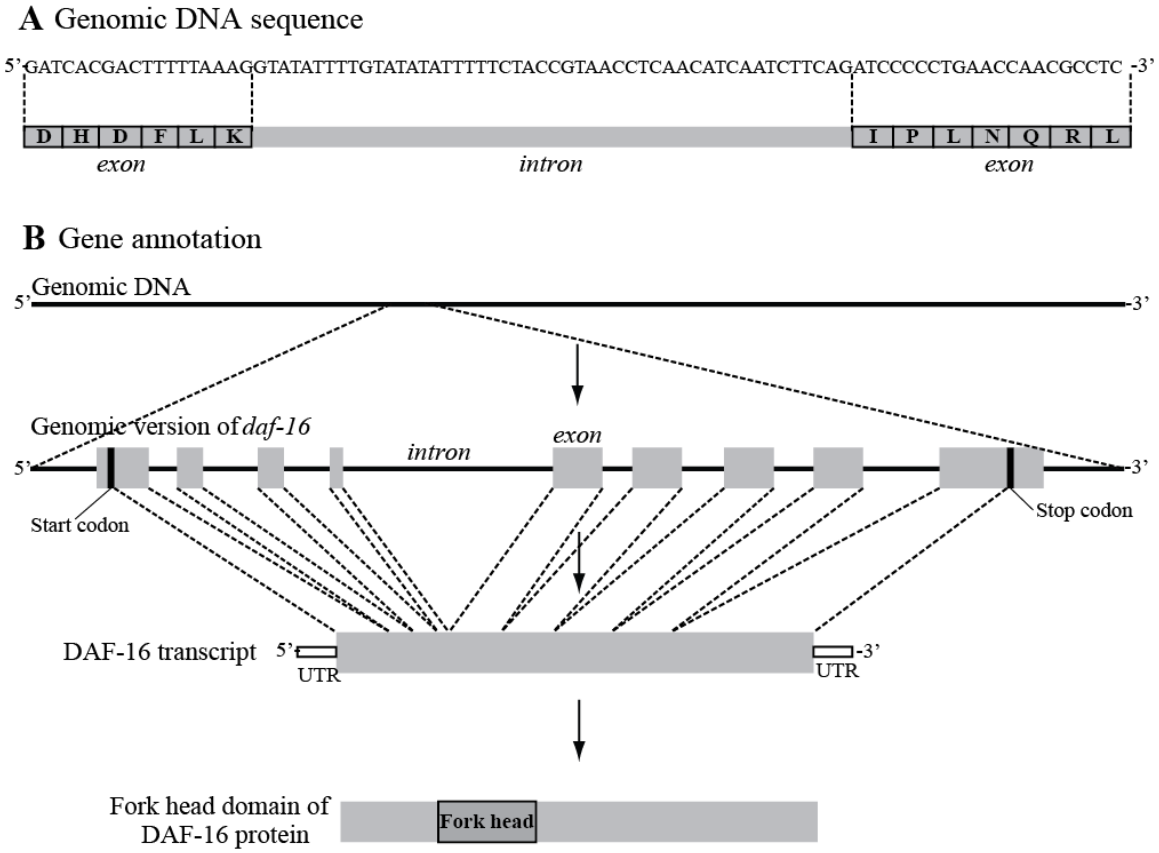

Figure 5.3 | Gene annotation. A) An overview of gene architecture, showing a stretch of DNA demarcated into protein-coding exons and non-coding introns. B) An example of gene annotation, proceeding from identification of genes from the genome either predicted *de novo* from gene-like patterns in the sequence or by comparison to known genes. In this case, the *daf-16* gene is shown with its introns and exons. Below that is a transcript from the *daf-16* gene, with the open reading frame (ORF) along with the upstream and downstream untranslated regions (UTR). The genome is also scanned for known protein domains; in this case, *daf-16* has a single protein domain, although other proteins may contain several domains.

Basic genomic metrics

The quality of a genome assembly can be assessed by metrics such as the total size of the genome relative to the fold coverage. This is estimated by dividing the total

number of assembled nucleotides by the genome size, which varies from 50–315 Mb for published and forthcoming nematode genomes (Table 5.1). For example, the *Ascaris suum* genome was sequenced with ~ 80–fold coverage, meaning that the 309 megabase genome was assembled from about 25 gigabases of sequence [12]. The GC content of the genome is usually reported, and varies between 27–48% among published and forthcoming nematode genomes (Table 5.1). Other commonly reported quality metrics of genomic assemblies address contiguity and completeness. One commonly used metric is the 'N50' value, which indicates that half of the genome is in contigs at least as large as that value. For instance, the N50 of the *A. suum* genome is 408 kb, meaning that half of the assembly is in contigs at least 408 kb in length [12]. Also important is the number of predicted protein coding genes, ranging from 13,000–45,000 among published and forthcoming genomes (Table 5.1). There are several other genomic statistics that have become potentially useful in comparisons such as gene density, number of transfer RNAs, and the percentage of high copy repeated sequences in the genome [8, 17, 18].

Quality assessments of genomic assembly provide confidence and a framework for interpreting subsequent analyses while other genomic metrics provide more information about the biological content of the genome. For instance, all known metazoan genomes require a certain number of tRNAs for codon recognition and for shuttling specific amino acids during translation, such that the number of tRNAs, tRNA pseudo-genes, and tRNA-derived repeats found in a genome assembly can serve as a rough estimate of completeness [42].

How protein sequences are analyzed and what they reveal about your nematode of choice

Annotation of nematode whole genomic sequence is complicated by several factors, including the structural complexity of introns, alternative RNA splicing, variable gene density, transplicing, and the presence of operons. Fortunately, annotation efforts on novel nematode species can leverage the excellent annotation of the *C. elegans* genome. These annotations are carefully curated and maintained in WormBase (www.wormbase.org), an expandable model for genome curation and annotation that already includes many available nematode genomes including *Ascaris suum, Brugia malayi, Bursaphelenchus xylophilus, Meloidogyne hapla, Meloidogyne incognita,* and many others. WormBase, with its established infrastructure and fulltime maintenance could serve as a repository for all nematode genomes and subsequent annotation [9]. As more genomes are sequenced and annotated, it has become clear that the availability of transcriptome data (e.g., RNA-seq; see above) is paramount for more accurate and comprehensive gene predictions, as well as elucidating biological function. While RNA requires more careful handling to avoid degradation, the reverse transcribed cDNA can be sequenced in the exact same manner as genomic DNA and for a similar cost.

While the specific details of annotation for each nematode genome differ, a general approach to protein analysis involves the following: identification of the protein-coding gene set, characterization by protein domain analysis and comparison to other protein databases, and comparative analysis with other nematodes and beyond. The identification of protein-coding genes is done using one or multiple gene prediction software packages, which generate *ab initio* predictions using machine-learning methods

such as hidden Markov models to identify open reading frames indicative of protein coding genes. The accuracy of these predictions can be improved by training the prediction software on experimental datasets such as ESTs, cDNA, protein similarity matches, and RNA-seq datasets. In particular, RNA-seq data can be used to partially or fully confirm gene-finder predictions [12, 17, 23, 43]. While computationally intensive, gene finding requires fewer resources than assembly.

As part of the annotation process, genes and proteins of the newly sequenced genome are evaluated by comparison to previously annotated genes and proteins from databases and genomes. Such evaluations identify putative homologous genes and proteins by sequence similarity. Homologous genes can be subdivided into orthologs and paralogs, depending on their history [44]. Orthologs are homologous sequences in different species that descended from a common ancestral gene during speciation, such that the ortholog of a gene in one species is the gene in the second species that shares decent from a common ancestral gene and is uniquely closely related to the gene in the first species. For example, the last common ancestor of *Pristionchus pacificus* and *C. elegans* may have possessed only one copy of the *daf-16* gene, which encodes a transcription factor in the insulin/IGF-1-mediated signaling pathway, and each of these extant species has one copy of *daf-16*, making these genes *daf-16* orthologs [8, 20, 45–47] (Figure 5.4A). We make this inference about *C. elegans* and *P. pacificus* knowing that both of these species as well as an outgroup taxon (in this case *A. suum*) all only have one copy of *daf-16*.

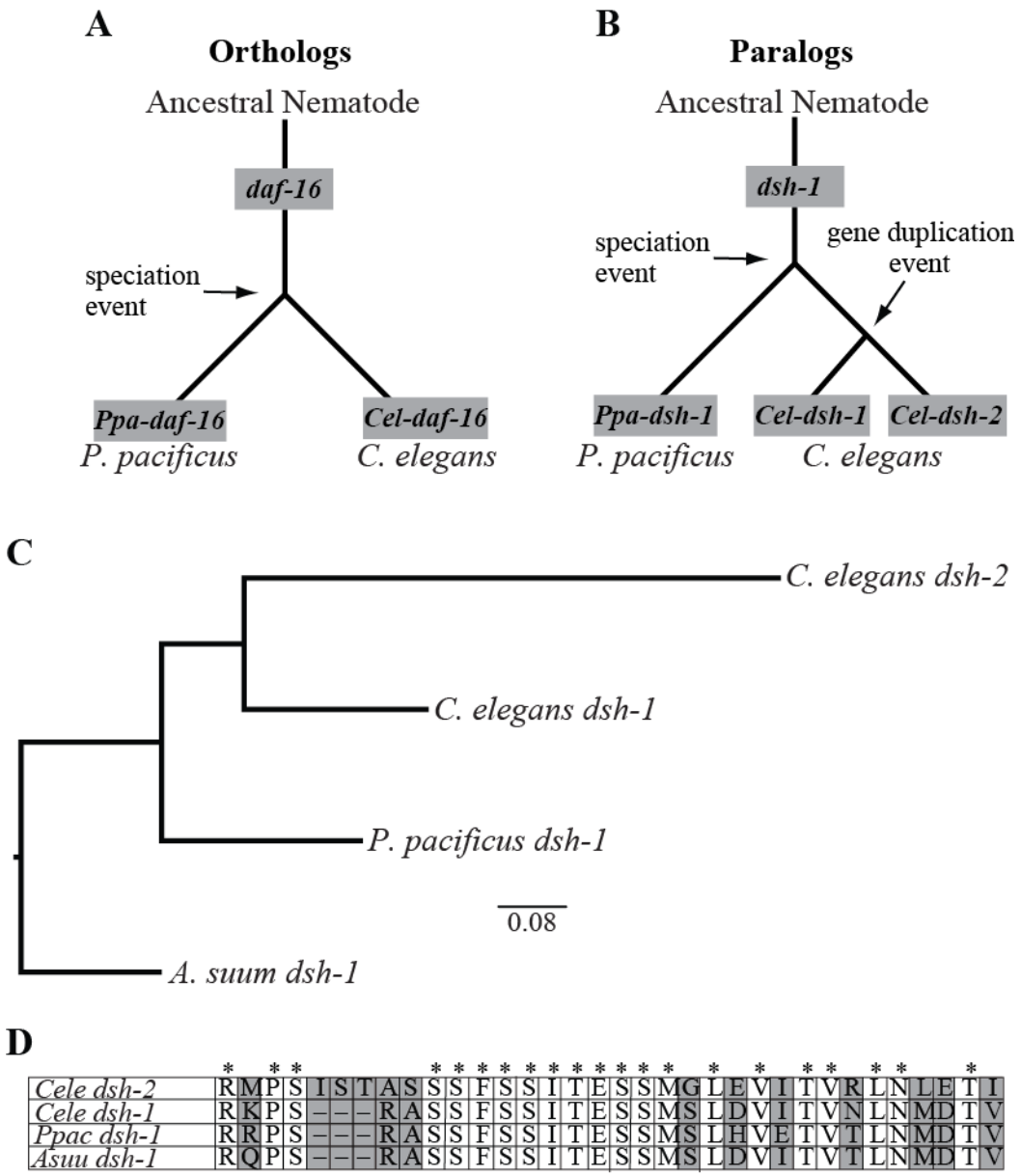

Figure 5.4 | Distinction between orthologous and paralogous genes. A) Orthologs are homologous sequences in different species that descend from a common ancestral gene during speciation. The *daf-16* gene in an ancestral nematode was conserved in both extant lineages resulting from a speciation event that lead to *P. pacificus* and *C. elegans*. *Ppa-daf-16* is the conserved *daf-16* gene in *P. pacificus* and *Cel-daf-16* is the conserved gene in *C. elegans*. B) Paralogs are homologous sequences within a species, having arisen by gene duplication or similar event. While *Ppa-dsh-1* and *Cel-dsh-1* are orthologs in this example, having both been conserved

from the same parental *dsh-1* copy, *Cel-dsh-1* and *Cel-dsh-2* are paralogs, having been duplicated within *C. elegans*. C) Neighbor-joining tree generated from gene comparisons of *dsh-1* homologs between *A. suum*, *P. pacificus*, and *C. elegans*, providing evidence that *Cel-dsh-1* is the ortholog of *Ppa-dsh-1* while *Cel-dsh-1* and *Cel-dsh-2* are paralogs. The small bar at the bottom center of the tree shows the approximate distance equal to 0.08 nucleotide changes. *A. suum* was used as the outgroup taxon (see Figure 2). The tree was made from an alignment of the full proteins in MUSCLE and subsequently analyzed using default parameters of the 'Dnadist' and 'Neighbor' programs from PHYLIP 3.68 software package [48, 49]. D) A 34 amino acid window of the protein alignment from which the tree in part C was generated. It shows the sequence conservation of the *dsh-1* orthologs and the subsequent divergence of *Cel-dsh-2* from the other sequences. Areas of sequence divergence are highlighted in grey while asterisks (*) indicate conserved amino acid identity across all four genes. Hyphens (-) indicate gaps in the alignment, likely the result of insertion/deletion events. *A. suum* serves as the outgroup taxon.

Paralogs are homologous sequences within a species, having arisen by gene duplication. Paralogs are thought *a priori* to share similar function, but this may not always be the case, as gene duplication and subsequent modification is thought to be the major way organisms evolve genes with novel functions [50]. For example, *P. pacificus* contains a single copy of the gene *dsh-1*, which encodes a signaling protein involved in embryogenesis, while *C. elegans* has two paralagous copies of the dishevelled gene, *dsh-1* and *dsh-2*. Relative to the outgroup *A. suum*, there appears to have been a duplication event in the *C. elegans* lineage since it diverged from *P. pacificus*; the last common ancestor of *P. pacificus* and *C. elegans* likely also possessed a single copy of this gene (Figure 5.4B) [47, 50, 51]. Based on higher sequence conservation with the sole *P. pacificus* protein, only *Cel-dsh-1* is considered to be a genuine ortholog of *Ppa-dsh-1*,

though experimental confirmation of conserved function would validate this inference (Figure 5.4C–D).

Once a gene set has been identified, putative functions are ascribed by database searching and similarity comparisons of the proteins from the new genome to those with known function. Commonly used databases include the NCBI BLAST database, the EMBL-EBI InterProScan, Pfam, and Gene Ontology databases [43, 52–55]. This initial assignment of protein function is based on the assumption of homology by sequence or domain similarity. In essence, the proteome (the full complement of protein coding genes) that results from whole genome sequencing and annotation has functions ascribed to its individual protein-coding sequences by comparing them to a number of different databases in search of sequence or domain similarity [20]. When a protein sequence from the genomic dataset has the highest degree of similarity to one sequence in another genome, it is *a priori* assumed to be homologous or to be derived from shared ancestry. The protein is further inferred to have similar function. In molecular phylogeny, homology infers shared ancestry. One important caveat of identifying homologs by sequence similarity is that it is not uncommon for two proteins to share functional similarity without shared ancestry, as a result of convergent evolution [47, 50, 56]. For example, *Heterorhabditis* and *Steinernema* nematodes utilize a specific type of insect parasitism and are known as entomopathogenic nematodes (EPNs), a characteristic they share not through ancestry but convergent evolution [34]. A notable molecular example is the convergent evolution of nearly identical antifreeze proteins in both Antarctic notothenioid fishes and Arctic cod, which show remarkable sequence and functional similarity that is due to evolutionary convergence rather than shared ancestry [57].

Another nematode example of convergence is the hermaphroditism of *C. elegans* and *C. briggsae*, which though outwardly similar as self-fertile hermaphrodites, have different molecular mechanisms for achieving this mode of reproduction [58]. The opposite caveat is also true; proteins of shared ancestry do not necessarily share similar function [59].

Orthologous gene associations across multiple genomes can provide powerful evolutionary insights into biological functions of individual genes as well as the evolution of species. They can be used to identify conserved genes, as in the case of pan-nematode genes or clade-specific genes. The identification of widely conserved or more specific genes serves as the basis for designing molecular diagnostic tools and elucidating the relationships between species. Multigene analyses from EST datasets have previously been successfully used to inform nematode phylogeny, and additional whole genome sequencing could identify new diagnostic markers to overcome sequencing identification difficulties and lack of phylogenetic resolution in some vexing taxa such as the tylenchids [60, 61]. Furthermore, such comparisons can be used in pursuit of non-conserved taxon-specific genes, which may reveal something about the particular biology and adaptations of individual species. For example, Kikuchi et al. (2011), in conjunction with publishing the *Bursaphelenchus xylophilus* genome included an orthology analysis across 10 nematode genomes. Although the genes shared across the 10 species did not fit an obvious phylogenetic pattern, the comparison revealed several gene families that are broadly conserved as well as small groups of genes shared between pairs or groups of nematodes that may be involved in the ecologies of those species. For example, 144 genes are shared exclusively between *P. pacificus* and *B. xylophilus* [25]. These nematodes occupy different ecological niches (one is necromenic and the other is a

migratory endoparasite of plants), but they both share a close association with insects during their lifecycle. Kikuchi *et al.* (2011) suggest that these genes are candidates for being involved in that association. The case for such a conclusion would be stronger if genome comparisons could show that the last common ancestor of both species also shared an association with insects.

Orthology analyses can also be used to explore the conservation of important biological pathways, such as sex determination, dauer formation, or the RNAi pathway. Because of the extent of detailed genetic exploration in *C. elegans*, a common starting place is to identify pathways of interest in *C. elegans* and search for their orthologs in another nematode of interest, though these results should be interpreted conservatively. For example, the RNAi pathway in *C. elegans* has been well-studied and found to be quite complex, with at least 77 genes known to be involved in core aspects of the process [33]. As a powerful reverse genetics technique, RNAi is a commonly examined pathway in newly sequenced genomes and has been developed as an experimental tool in both plant- and animal-parasitic nematodes including *Globodera pallida*, *Heterordera glycines*, *M. incognita*, and *B. malayi* [62–64]. It may even have practical utility in agriculture in controlling plant-parasitic nematodes or at least increasing plant resistance [65, 66]. How many of the 77 known RNAi effector genes are absolutely necessary for RNAi in general and how many are part of the specific mechanism of RNAi in *C. elegans*? For instance, *sid-1* is necessary for systemic RNAi in *C. elegans*, but systemic RNAi has been reported in several other species that do not seem to contain an identifiable homolog of *sid-1*, including *B. malayi*, *Globodera* and *Meloidogyne* spp., *Pristionchus pacificus*, and *Panagrolaimus superbus* [62, 63, 67–70]. The successful application of experimental

RNAi in species that are apparently missing some genes required for systemic RNAi in *C. elegans* implies that either these genes are rapidly evolving or have only become necessary in *C. elegans,* or that an alternate pathway exists [18, 33, 62]. Although RNAi has been shown to work in a number of both plant- and animal- parasitic nematodes, it is thought that culturability and the feasibility of maintaining non-stressful culturing conditions may better explain RNAi competencies than the disparity of RNAi effector genes across taxa [33]. As more species are added to these types of genomic analyses and genetic experimentation in non-model systems continues to grow, our understanding of these processes and which parts are conserved, derived, or rapidly evolving will become more clear.

Operons

One striking feature of nematode genomes studied thus far is the presence of operons. Though originally thought to be a genomic feature unique to prokaryotes, operons have been found in nematodes as well as some ascidians and fruit flies [71]. Bacterial operons comprise 2 or more genes that are transcribed to form a single mRNA transcript (Figure 5.5). In nematodes, multiple genes are transcribed into a single primary transcript, which is then processed into separate mRNAs; through RNA-splicing events, a spliced leader is added to the 5' end of each downstream transcript in operon (Figure 5.5). In *C. elegans*, about 70% of mRNAs include a spliced leader, the majority of which (~ 55%) are of the SL1 type. These SL1 spliced leaders are typically either from non-operonic transcripts or are from the first gene in an operon (Figure 5.5) [72]. Downstream transcripts from within an operon each have an SL2 leader [72]. Operons can be inferred

from the genome by the presence of very closely spaced genes in the same orientation in the genome and from the presence of SL2 spliced leaders. Apparent operons have been identified in all published nematode genomes with the exception of *Trichinella spiralis*, a highly unusual nematode, quite distantly related to all other sequenced nematodes and one of the world's largest intracellular parasites (Figure 5.2) [24, 73]. Although *T. spiralis* is missing both canonical nematode *trans*-spliced leaders, SL1 and SL2, the presence of a number of other distinct spliced leader sequences leaves open the possibility that this species does contain operons. Additional nematode genomic data, especially from taxa in Enoplia, Dorylaimia, and basal clades of Chromadoria, may reveal the untold story of operon evolution among nematodes (Figure 5.2). Operons are thought to have evolved in nematodes to facilitate transitions from arrested development to rapid growth [74].

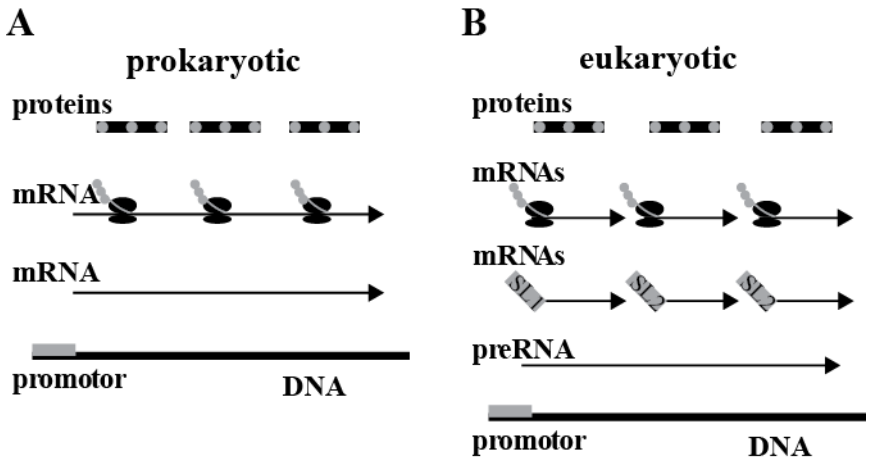

Figure 5.5 | Prokaryotic and eukaryotic operons. A) The prokaryotic operon model is that polycistronic mRNA is produced from a single promoter, containing several genes in the same transcript. Each protein is translated from a different location of a single mRNA. B) Eukaryotic operons also produce a single preRNA transcript with multiple genes, but are then processed to

form mRNAs for each individual protein. As part of the processing each mRNA has a splice leader added to the 5' end of the transcript. These mRNAs are then translated. Usually the first gene from the preRNA in a nematode operon transcript is spliced with the SL1 splice leader attached while all other downstream genes have the SL2 splice leader attached.

Genomes and ecology

The first report of a nematode genome focused on the sequencing methodology, the development of physical and genetic maps, assembly, and annotation, as well as a comparison of the genome to prokaryotes and yeast [8, 14]. This comparison revealed that *C. elegans* has an unusually high number of nuclear hormone receptor proteins (NHRs), prompting researchers to propose that NHRs were perhaps important in the evolution of multicellularity [8]. Though originally thought to be normal among nematodes, it is now known that even among close relatives, *C. elegans* is an outlier in terms of its number of NHRs and G protein-coupled receptors (GPCRs) and in these respects is not an archetypical nematode [6, 75]. The anomalously high number of NHRs and GPCRs in the *C. elegans* genome was found by examining the top 20 most prevalent protein domains in the genome. Such comparisons of gene and domain prevalence among species may reveal important differences in the genome that ultimately underlie differences in the evolution, ecology, and lifestyles of nematodes. In this way, comparative genome analyses will serve as a tool for testing hypotheses about the ecology and evolution of related species; and the resolving power of such comparisons will increase with the addition of more sequenced taxa.

The sequencing of *C. briggsae* greatly enhanced our understanding of the *C. elegans* genome by providing strong evidence for 1,300 previously unidentified genes,

thus demonstrating how sequencing closely related species can enhance the annotation of genomes [17]. Analysis of repeat regions revealed that *C. elegans* and *C. briggsae* have undergone rapid evolutionary turnover at the sequence level, providing evidence for a more recent divergence of these two nematodes compared to the evolutionary split between human and mouse lineages (~ 40 million years ago for the nematodes and ~ 75 million years ago for mouse/human). Similarly, the amino acid identity revealed between putative orthologs (~ 80% for *C. briggsae*/*C. elegans* and ~ 78.5% for mouse/human) supports this conclusion [17, 76].

As sequencing technology has advanced and costs have dropped, additional nematode genomes have been sequenced, including close relatives of *C. elegans* (*C. angaria, C. brenneri*, *C. japonica*, *and C. remanei*) and a handful of economically important parasites such as *Bursaphelenchus xylophilus*, *Meloidogyne incognita*, and *M. hapla* [12, 18, 19, 21–25]. One of the rationales for vertebrate-parasitic nematode sequencing projects (*B. malayi, A. suum*, and *T. spiralis*) was the identification of candidate genes to target pharmacologically [12, 18, 24]. This is a particularly important avenue of research given the large number of humans affected by nematode diseases and our current reliance on a small pool of drugs, whose effectiveness is at risk due to increasing resistance [77]. In addition to identifying new drug targets, these genomic analyses identified genes likely to be involved in the vertebrate-parasitic lifestyle, or perhaps parasitism in general. The abundance and diversity of secreted proteases and protease inhibitors in these genomes was an interesting result and has produced a long list of genes that are candidates to be involved in invasion of host tissues and degradation or evasion of host immune responses. The *B. malayi* genome's lack of key metabolic

enzymes provided evidence for this nematode's reliance on host- or *Wolbachia*-supplied molecules for purine, riboflavin, and heme biosynthesis [18]. Due to the basal position of *T. spiralis* in Dorylaimia (Figure 5.2), its genome was compared to all other available nematode genomes to identify pan-Nematoda-specific conservation. The resulting list of genes and proteins may have fundamental importance in all nematodes and points to potential targets for control of parasitic nematodes throughout the phylum [24]. Because of the highly specific and derived lifestyle of *T. spiralis*, which is an intracellular parasite, it is likely that examination of additional basal taxa will improve and solidify a pan-Nematoda candidate gene list, which, in addition to providing potential pharmacological targets could be used to inform deeper level phylogenetic studies.

Root-knot nematodes are among the most agriculturally devastating plant pathogens known in any phylum [19, 78]. This motivated the sequencing of *Meloidogyne incognita*, closely followed by *Meloidogyne hapla* [19, 21]. These genomes have provided intriguing insights into the adaptive strategies used by metazoans to circumvent immunity and successfully parasitize plants [19, 21]. They also provided evidence to support the long-suspected role of horizontal gene transfer (HGT) in the evolution of plant parasitism [79, 80]. Both of these parasites seem to have benefitted from the acquisition of plant cell wall-degrading enzymes that appear bacterial in origin. The idea that nematodes can acquire and utilize such enzymes in a cross-kingdom way was further bolstered by similar findings from genomic analyses of the mycophagous plant parasite *B. xylophilus* and the necromenic species *P. pacificus* [20, 25]. Recent follow-up work on HGT in multiple *Pristionchus* and related species utilized genome, transcriptome, and EST data sets, and revealed functional laterally acquired cellulase genes in several

diplogastrid species, notable turnover of cellulase genes inferred from elevated gene birth and death rates, and showed evidence for selective forces working on individual cellulase genes with a high degree of specificity [81]. Moreover, some cellulases found in *B. xylophilus* have not been found in any other nematode and appear fungal in origin, providing evidence that, if these genes are the result of HGT and not the independently arising result of convergent evolution, nematodes may not be limited to bacteria as sources of adaptational armament [25]. The evidence for HGT in multiple distantly related nematodes (*Bursaphelenchus*, Koerneria, *Meloidogyne*, and *Pristionchus*) suggests that this mode of gene acquisition may play a broadly significant role in nematode adaptation and evolution (Figure 5.2).

One clear theme that has emerged from genomic comparisons is that there may not be an archetypal nematode [6, 75]. For example, the massive expansions in GPCRs and NHRs reported in *C. elegans* are thus far not replicated in the genomes of any other sequenced nematodes, and likely play a significant role in *C. elegans*' natural ecology, which has only recently been explored through modern investigation [82–84]. As more nematode species are fully sequenced, it is becoming clear that the ecology and specific biology of each species will become increasingly valuable in the interpretation and use of these genomes. While earlier reports of nematode genomes focused heavily on sequencing methodologies and the technical details of gene prediction and annotation, more recent studies have highlighted genomes in the context of nematode ecology and evolution; this trend is likely to continue. For instance, *P. pacificus* is an omnivorous feeder, necromenic but not parasitic. It associates with arthropods and waits for them to die, feasting on the microbial and fungal bloom resulting from the arthropod host's death

[82, 85]. A broad view of the *P. pacificus* genome reveals expansions in protein families playing key roles in stress tolerance and the metabolism of xenobiotics (foreign chemical compounds; e.g., host defense molecules) [20]. Tolerance to low oxygen concentrations and toxic host enzymes as well as complex metabolic pathways and other morphological adaptations were predicted to assist this nematode in its lifestyle, but prior to its genome being sequenced the molecular architecture of these adaptations could only be speculative [20]. The genetic underpinnings of necromeny in *P. pacificus* and its adaptation to this particular niche have been revealed through its genome. These findings lead to additional genomically generated hypotheses and sow fertile ground for future experimentation.

Ecological genomics is a burgeoning field aimed at understanding the genetic mechanisms that underlie organismal responses and adaptations to their natural environments [86]. Model organisms, often chosen for ease of culture and a host of other traits that favor laboratory growth and experimentation, usually lack the extensive ecological context and framework that has been painstakingly built for many non-model systems. In contrast, many organisms used in ecological studies do not have the extensive experimental tool development (e.g., transformation and RNAi) or genetic pathways and interactions mapped out as in model systems. The time is ripe for dramatic expansion of ecological studies using model systems and genomic/transcriptomic sequencing and accompanying tool development to be done in favored ecological systems [87]. Nematodes are in a superb position to see progress in both areas, with several well-developed model systems being explored from an ecological context [82–84, 88, 89] and for nematode species for which archives of ecological data have been accumulated to be scrutinized from a genomic context [90, 91].

Entomopathogenic nematodes as an example of question-driven genomics

Nematode genomics, now highlighting specific aspects of organismal biology, life history traits, and ecology and evolution, provides opportunity for researchers to utilize the powerful broad view of sequencing to learn more about their nematode of choice. As an illustrative example of ecological genomics and what could be accomplished for every niche occupied by nematodes, we conclude by discussing some of the interesting genomic insights that can be gleaned from examining the forthcoming entomopathogenic nematode genomes.

EPNs occupy an interesting niche somewhere between parasitoids and pathogens, utilizing insect-pathogenic bacteria to facilitate their form of parasitism, acting as a vector for the bacteria and, working together as a complex, the nematode and bacteria rapidly kill their host [92, 93]. This very specific form of parasitism seems to have arisen at least twice among nematodes, in *Heterorhabditidae* and *Steinernematidae*, which are not closely related. The genomic sequencing of heterorhabditid and steinernematid nematodes will provide the framework for a genetic comparison of the evolution of entomopathogeny in these lineages [87]. In contrast to the vertebrate- and plant-parasitic nematode genome studies, which compare organisms that obtain resources by different means, the intra-guild comparisons of EPN genomes will focus on species that exploit the same kind of environmental resources in similar ways [94, 95]. A genomic comparison of EPNs from multiple genera has the advantage of decades of ecological research and will

increase our understanding of adaptation and convergent evolution in addition to revealing just how similar or different this niche exploitation is at the genetic level.

EPNs have rapidly become models for studying parasitism and mutualism. The genetic components of their association with symbiotic bacteria have been heavily studied from the bacterial side, but largely neglected in terms of the nematode's contribution [90, 96]. Genome-wide expression analysis against the backdrop of the genomic sequence could shed light on what, if any, contribution is made by the nematodes to symbiosis. Within *Steinernema*, there are more than 60 described species [97–105]. Though only a handful of these have been tested, the host-range and specificity of insects they can infect is diverse and varied. A striking example is *S. carpocapsae*, which is the most heavily studied steinernematid. With an extremely broad host range, *S. carpocapsae* is capable of infecting more than 250 species of insects across 10 orders, although some infections were only demonstrated under laboratory conditions [106]. Closely related to *S. carpocapsae* is *S. scapterisci*, which is known to have a much narrower host range and seems to be a cricket specialist [107, 108]. The wide view afforded by protein family abundances revealed by genomes will provide testable hypotheses about the breadth of specific of EPNs' host-range and the specificity of some EPNs for certain insect hosts, beyond what is currently known.

EPN research has also seen recent developments in the neuronal basis of behavior and the molecular mechanisms underlying host tissue invasion and death [91, 109]. Understanding protein domain abundance against this backdrop will likely hone existing hypotheses and direct future experimentation, leading to a deepening of our knowledge in both of these areas of research. Along with the broad overview on the architecture of

parasitism, it is anticipated that EPN genomes will provide insights to the above mentioned and other aspects of EPN ecology. A hopeful expectation of most new nematode genomes is that they will pave the way for techniques such as transformation and RNAi to be used in experimentally testing the genomically generated hypotheses, as exemplified with *P. pacificus* [20, 68].

Conclusion

Many new nematode genomic sequencing projects are underway, and improving technologies means still more will become feasible and affordable. These widening horizons are generating a need for more nematodes to be cultured and have their DNA harvested. More importantly, it opens the door for collaborations between genomicists and nematologists. We expect that fruitful collaborations will entail far more than merely providing material and could include various aspects such as (a) knowledge of the ecological background and candidate pathways or biological phenomena to explore within the sequence, (b) phylogenetic knowledge of sister taxa or associated nematodes for comparison or particularly informative developmental stages for transcript analysis, and (c) interesting morphological features that remain to be genetically explored. We urge the members of the Society of Nematologists to utilize their expertise and the wealth of their collective ecological knowledge to contribute to sequencing efforts and to adopt genomics into the toolkit of nematology. As nematology stands at the precipice of genomic grandeur, with 959 nematode genomes planned (a number chosen to reference the 959 somatic cells of *C. elegans* [26]), we will soon be suffused with genomic data,

offering the potential to discover long-sought answers to the biology, ecology, and evolution of genomes, and promising in turn to raise many more new questions.

References

1. Yeates, G.W. (2004). Ecological and behavioral adaptations. In Nematode Behavior, R. Gaugler and A.L. Bilgramis, eds. (Cambridge: CABI), pp. 1–18.

2. De Ley, P., and Blaxter, M. (2002). Systematic position and phylogeny. In The Biology of Nematodes, D. Lee, ed. (London: Taylor & Francis), pp. 1–30.

3. De Ley, P., and Blaxter, M. (2004). A new system for Nematoda: combining morphological characters with molecular trees, and translating clades into ranks and taxa. In Nematology Monographs and Perspectives, Volume 2, R. Cook and D.J. Hunt, eds. (Leiden: E.J. Brill), pp. 633–653.

4. Holterman, M., van der Wurff, A., van den Elsen, S., van Megen, H., Bongers, T., Holovachov, O., Bakker, J., and Helder, J. (2006). Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. Molecular Biology and Evolution *23*, 1792–1800.

5. Lin, M.F., Carlson, J.W., Crosby, M.A., Matthews, B.B., Yu, C., Park, S., Wan, K.H., Schroeder, A.J., Gramates, L.S., St. Pierre, S.E., et al. (2007). Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. Genome Res. *17*, 1823–1836.

6. Blaxter, M. (2011). Nematodes: The worm and its relatives. PLoS Biology *9*, 1–9.

7. Burgess, D.J. (2011). Comparative genomics: Mammalian alignments reveal human functional elements. Nat. Rev. Genet. *12*, 806–807.

8. Consortium (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science *282*, 2012–2018.

9. Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R.H., et al. (2010). WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. *38*, D463–D467.

10. Horvitz, H.R., and Sternberg, P.W. (1982). Nematode postembryonic cell lineages. J. Nematol. *14*, 240–248.

11. Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics *9*, 387–402.

12. Jex, A.R., Liu, S., Li, B., Young, N.D., Ross, S.H., Li, Y., Yang, L., Zeng, N., Xu, X., Xiong, Z., et al. (2011). *Ascaris suum* draft genome. Nature *479*, 529–533.

13. Werner, J.J., Zhou, D., Caporaso, J.G., Knight, R., and Angenent, L.T. (2011). Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. ISME J, doi: 10.1038/ismej.2011.1186.

14. Consortium (1999). How the worm was won. Trends Genet *15*, 51–58.

15. Collins, F.S., Drumm, M.L., Cole, J.L., Lockwood, W.K., Van de Woude, G.F., and Iannuzzi, M.C. (1987). Construction of a general human chromosome jumping library, with application to cystic fibrosis. Science *235*, 1046–1049.

16. Richards, J.E., Gilliam, T.C., Cole, J.L., Drumm, M.L., Wasmuth, J.J., Gusella, J.F., and Collins, F.S. (1988). Chromosome jumping from D4S10 (G8) toward the Huntington disease gene. Proc. Natl. Acad. Sci. U. S. A. *85*, 6437–6441.

17. Stein, L.D., Bao, Z.R., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N.S., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. (2003). The genome sequence of Caenorhabditis briggsae: A platform for comparative genomics. PLoS Biology *1*, e45. doi:10.1371/journal.pbio.0000045.

18. Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J.E., Delcher, A.L., Guiliano, D.B., Miranda-Saavedra, M., et al. (2007). Draft genome of the filarial nematode parasite *Brugia malayi*. Science *317*, 1756–1760.

19. Abad, P., Gouzy, J., Aury, J.M., Castagnone-Sereno, P., Danchin, E.G.J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., et al. (2008). Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nature Biotechnology *26*, 909–915.

20. Dieterich, C., Clifton, S.W., Schuster, L.N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., Fulton, L., Fulton, R., Godfrey, J., Minx, P., et al. (2008). The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. Nature Genet. *40*, 1193–1198.

21. Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., et al. (2008). Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. Proc. Natl. Acad. Sci. U. S. A. *105*, 14802–14807.

22. Barrière, A., Yang, S.-P., Pekarek, E., Thomas, C.G., Haag, E.S., and Ruvinsky, I. (2009). Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. Genome Res. *19*, 470–480.

23. Mortazavi, A., Schwarz, E.M., Williams, B.A., Schaeffer, L., Antoshechkin, I., Wold, B., and Sternberg, P.W. (2010). Scaffolding a *Caenorhabditis* nematode genome with RNA-seq. Genome Res. *20*, 1740–1747.

24. Mitreva, M., Jasmer, D.P., Zarlenga, D.S., Wang, Z., Abubucker, S., Martin, J., Taylor, C.M., Yin, Y., Fulton, L., Minx, P., et al. (2011). The draft genome of the parasitic nematode *Trichinella spiralis*. Nature Genet. *43*, 228–236.

25. Kikuchi, T., Cotton, J.A., Dalzell, J.J., Hasegawa, K., Kanzaki, N., McVeigh, P., Takanashi, T., Tsai, I.J., Assefa, S.A., Cock, P.J.A., et al. (2011). Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. PLoS Pathogens *7*, e1002219.

26. Kumar, S., Schiffer, P.H., and Blaxter, M. (2012). 959 nematode genomes: a semantic wiki for coordinating sequencing projects. Nucleic Acids Res. *40*, D1295–D1300.

27. Lambshead (1993). Recent developments in marine benthic biodiversity research. Oceanis *19*, 5–24.

28. Lambshead, P.J. (2004). Marine nematode biodiversity. In Nematode morphology, physiology, and ecology, Z.X. Chen, Y. Chen, S.Y. Chen and D.W. Dickson, eds. (CABI), pp. 4554–4558.

29. Hoda, M. (2011). Phylum Nematoda Cobb 1932. In Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness., Z.-Q. Zhang, ed., pp. 63–95.

30. van Megen, H., van Den Elsen, S., Holterman, M., Karssen, G., Mooyman, P., Bongers, T., Holovachov, O., Bakker, J., and Helder, J. (2009). A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology *11*, 927–950.

31. Schierenberg, E. (2005). Unusual cleavage and gastrulation in a freshwater nematode: developmental and phylogenetic implications. Development Genes and Evolution *215*, 103–108.

32. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

33. Dalzell, J.J., McVeigh, P., Warnock, N.D., Mitreva, M., Bird, D.M., Abad, P., Fleming, C.C., Day, T.A., Mousley, A., Marks, N.J., et al. (2011). RNAi effector diversity in nematodes. PLoS Neglected Tropical Diseases *5*, e1176.

34. Adams, B.J., Peat, S.M., and Dillman, A.R. (2007). Phylogeny and evolution. In Entomopathogenic nematodes: Systematics, phylogeny, and bacterial symbionts., Volume 5, K.B. Nguyen and D.J. Hunt, eds. (Leiden-Boston: Brill), pp. 693–733.

35. Williams, B.D., Schrank, B., Huynh, C., Shownkeen, R., and Waterston, R.H. (1992). A genetic mapping system in Caenorhabditis elegans based on polymorphic sequence-tagged sites. Genetics *131*, 609–624.

36. Jones, K.L., Todd, T.C., and Herman, M.A. (2006). Development of taxon-specific markers for high-throughput screening of microbial-feeding nematodes. Molecular Ecology Notes *6*, 712–714.

37. Zheng, Z., Advani, A., Melefors, O., Glavas, S., Nordstrom, H., Ye, W., Engstrand, L., and Andersson, A.F. (2011). Titration-free 454 sequencing using Y adapters. Nature protocols *6*, 1367–1376.

38. Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

39. Li, R., Yu, C., Li, L., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.

40. Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

41. Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. *32*, W309–W312.

42. Bermudez-Santana, C., Attolini, C.S., Kirsten, T., Engelhardt, J., Prohaska, S.J., Steigele, S., and Stadler, P.F. (2010). Genomic organization of eukaryotic tRNAs. BMC genomics *11*, 270.

43. Pevsner, J. (2009). Protein Analysis and Proteomics. In Bioinformatics and Functional Genomics. (Hoboken: John Wiley & Sons), pp. 379–416.

44. Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Systematic Zoology *19*, 99–113.

45. Riddle, D.L., Swanson, M.M., and Albert, P.S. (1981). Interacting genes in nematode dauer larva formation. Nature *290*, 668–671.

46. Sonnhammer, E.L., and Durbin, R. (1997). Analysis of protein domain families in *Caenorhabditis elegans*. Genomics *46*, 200–216.

47. Pevsner, J. (2009). Pairwise sequence alignment. In Bioinformatics and Functional Genomics. (Hoboken: John Wiley & Sons), pp. 47–97.

48. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

49. Felsenstein, J. (1989). PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics *5*, 164–166.

50. Graur, D., and Li, W.-H. (2000). Gene duplication, exon shuffling, and concerted evolution. In Fundamentals of Molecular Evolution, D. Graur and W.-H. Li, eds. (Sunderland, MA: Sinauer Associates), pp. 249–322.

51. Eisenmann, D. (2005). Wnt signaling. WormBook.

52. Zbodnov, E.M., and Apweiler, R. (2001). InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.

53. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths–Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *1*, D138–D141.

54. Ye, J., McGinnis, S., and Madden, T.L. (2006). BLAST: improvements for better sequence analysis. Nucleic Acids Res. *34*, W6–W9.

55. Thomas, P.D., Mi, H., and Lewis, S. (2007). Ontology annotation: Mapping genomic regions to biological function. Current Opinion in Chemical Biology *11*, 4–11.

56. Graur, D., and Li, W.-H. (2000). Genome evolution. In Fundamentals of Molecular Evolution, D. Graur and W.-H. Li, eds. (Sunderland, MA: Sinauer Associates), pp. 367–427.

57. Chen, L., DeVries, A.L., and Cheng, C.-H.C. (1997). Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fishes and Arctic cod. Proc. Natl. Acad. Sci. U. S. A. *94*, 3817–3822.

58. Hill, R.C., de Carvalho, C.E., Salogiannis, J., Schlager, B., Pilgrim, D., and Haag, E.S. (2006). Genetic flexibility in the convergent evolution of hermaphroditism in Caenorhabditis nematodes. Dev Cell *10*, 531–538.

59. Sangar, V., Blankenberg, D.J., Altman, N., and Lesk, A.M. (2007). Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics *8*.

60. Adams, B.J., Dillman, A.R., and Finlinson, C. (2009). Molecular taxonomy and phylogeny. In Root-knot Nematodes, R.N. Perry, M. Moens and J.L. Starr, eds. (CABI), pp. 119–138.

61. Scholl, E.H., and Bird, D.M. (2005). Resolving tylenchid evolutionary relationships through multiple gene analysis derived from EST data. Molecular phylogenetics and evolution *36*, 536–545.

62. Aboobaker, A.A., and Blaxter, M. (2003). Use of RNA interference to investigate gene function in the human filarial nematode parasite *Brugia malayi*. Molecular Biochemical Parasitology *129*, 41–51.

63. Dalzell, J.J., McMaster, S., Fleming, C.C., and Maule, A.G. (2010). Short interfering RNA-mediated gene silencing in Globodera pallida and Meloidogyne incognita infective stage juveniles. Int J Parasitol *40*, 91–100.

64. Urwin, P.E., Lilley, C.J., and Atkinson, H.J. (2002). Ingestion of double-stranded RNA by preparasitic juvenile cyst nematodes leads to RNA interference. Molecular plant-microbe interactions: MPMI *15*, 747–752.

65. Huang, G., Allen, R., Davis, E.L., Baum, T.J., and Hussey, R.S. (2006). Engineering broad root-knot resistance in transgenic plants by RNAi silencing of a conserved and essential root-knot nematode parasitism gene. Proc Natl Acad Sci U S A *103*, 14302–14306.

66. Yadav, B.C., Veluthambi, K., and Subramaniam, K. (2006). Host-generated double stranded RNA induces RNAi in plant-parasitic nematodes and protects the host from infection. Molecular and biochemical parasitology *148*, 219–222.

67. Shannon, A.J., Tyson, T., Dix, I., Boyd, J., and Burnell, A.M. (2008). Systemic RNAi mediated gene silencing in the anhydrobiotic nematode Panagrolaimus superbus. BMC Mol Biol *9*.

68. Cinkornpumin, J.K., and Hong, R.L. (2011). RNAi mediated gene knockdown and transgenesis by microinjection in the necromenic nematode *Pristionchus pacificus*. Journal of Visualized Experiments *56*, e3270.

69. Kimber, M.J., McKinney, S., McMaster, S., Day, T.A., Fleming, C.C., and Maule, A.G. (2007). flp gene disruption in a parasitic nematode reveals motor dysfunction and unusual neuronal sensitivity to RNA interference. FASEB journal: official publication of the Federation of American Societies for Experimental Biology *21*, 1233–1243.

70. Rosso, M.N., Jones, J.T., and Abad, P. (2009). RNAi and functional genomics in plant parasitic nematodes. Annu Rev Phytopathol *47*, 207–232.

71. Blumenthal, T. (2004). Operons in eukaryotes. Briefings in Functional Genomics and Proteomics *3*, 199–211.

72. Blumenthal, T. (2005). Trans-splicing and operons. In WormBook.

73. Despommier, D.D. (1990). *Trichinella spiralis*: The worm that would be virus. Parasitology Today *6*, 193–196.

74. Zaslaver, A., Baugh, L.R., and Sternberg, P.W. (2011). Metazoan operons accelerate recovery from growth-arrested states. Cell *145*, 981–992.

75. Blaxter, M. (1998). *Caenorhabditis elegans* is a nematode. Science *282*, 2041–2046.

76. Cutter, A.D. (2008). Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. Molecular Biology and Evolution *25*, 778–786.

77. Keiser, J., and Utzinger, J. (2010). The Drugs We Have and the Drugs We Need Against Major Helminth Infections. Adv Parasit *73*, 197–230.

78. Trudgill, D.L., and Blok, V.C. (2001). Apomictic, polyphagous rook-knot nematodes: exceptionally successful and damaging biotrophic root pathogens. Annual Review of Phytopathology *39*, 53–77.

79. Smant, G., Stokkermans, J.P.W.G., Yan, Y., De Boer, J.M., Baum, T.J., Wang, X., Hussey, R.S., Gommers, F.J., Henrissat, B., Davis, E.L., et al. (1998). Endogenous cellulases in animal: Isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. Proc. Natl. Acad. Sci. U. S. A. *95*, 4906–4911.

80. Popeijus, H., Overmars, H., Jones, J., Blok, V.C., Goverse, A., Helder, J., Schots, A., Bakker, J., and Smant, G. (2000). Degradation of plant cell walls by a nematode. Nature *406*, 36–37.

81. Mayer, W.E., Schuster, L.N., Bartelmes, G., Dieterich, C., and Sommer, R.J. (2011). Horizontal gene transfer of microbial cellulases into nematode genomes is associated with functional assimilation and gene turnover. BMC evolutionary biology *11*, 13.

82. Kiontke, K., and Sudhaus, W. (2006). Ecology of *Caenorhabditis* species. In WormBook.

83. Troemel, E.R., Félix, M.A., Whiteman, N.K., Barriere, A., and Ausubel, F.M. (2008). Microsporidia are natural intracellular parasites of the nematode *C. elegans*. PLoS Biol *6*, e309.

84. Félix, M.A., and Braendle, C. (2010). The natural history of *Caenorhabditis elegans*. Current Biology *20*, 965–969.

85. Herrmann, M., Mayer, W.E., and Sommer, R.J. (2006). Nematodes of the genus *Pristionchus* are closely associated with scarab beetles and the Colorado potato beetle in Western Europe. Zoology *109*, 96–108.

86. Ungerer, M.C., Johnson, L.C., and Herman, M.A. (2008). Ecological genomics: understanding gene and genome function in the natural environment. Heredity (Edinb) *100*, 178–183.

87. Elmer, K.R., and Meyer, A. (2011). Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends in ecology & evolution *26*, 298–306.

88. Mayer, M.G., and Sommer, R.J. (2011). Natural variation in Pristionchus pacificus dauer formation reveals cross-preference rather than self-preference of nematode dauer pheromones. Proceedings. Biological sciences / The Royal Society *278*, 2784–2790.

89. Rae, R., Riebesell, M., Dinkelacker, I., Wang, Q., Herrmann, M., Weller, A.M., Dieterich, C., and Sommer, R.J. (2008). Isolation of naturally associated bacteria of necromenic *Pristionchus* nematodes and fitness consequences. J. Exp. Biol. *211*, 1927–1936.

90. Ciche, T.A., and Sternberg, P.W. (2007). Postembryonic RNAi in Heterorhabditis bacteriophora: a nematode insect parasite and host for insect pathogenic symbionts. BMC developmental biology *7*, 101.

91. Hallem, E.A., Dillman, A.R., Hong, A.V., Zhang, Y., Yano, J.M., DeMarco, S.F., and Sternberg, P.W. (2011). A sensory code for host seeking in parasitic nematodes. Current Biology *21*, 377–383.

92. Kaya, H.K., and Gaugler, R. (1993). Entomopathogenic nematodes. Annu. Rev. Entomol. *38*, 181–206.

93. Dillman, A.R., Chaston, J.M., Adams, B.J., Ciche, T.A., Goodrich-Blair, H., Stock, S.P., and Sternberg, P.W. (2012). An entomopathogenic nematode by any other name. PLoS Pathogens *8*, e1002527.

94. Root, R.B. (1967). The niche exploitation pattern of the blue-gray gnatcatcher. Ecological Monographs *37*, 317–350.

95. Fauth, J.E., Bernardo, J., Camara, M., Resetarits, W.J., VanBuskirk, J., and McCollum, S.A. (1996). Simplifying the jargon of community ecology: A conceptual approach. Am Nat *147*, 282–286.

96. Chaston, J., and Goodrich-Blair, H. (2010). Common trends in mutualism revealed by model associations between invertebrates and bacteria. FEMS Microbiology Reviews *34*, 41–58.

97. Edgington, S., Buddie, A.G., Tymo, L.M., Hunt, D.J., Nguyen, K.B., France, A.I., Merino, L.M., and Moore, D. (2009). *Steinernema australe* n. sp. (Panagrolaimorpha: Steinernematidae) a new entomopathogenic nematode from Isla Magdalena, Chile. Nematology *11*, 699–717.

98. Khatri-Chhetri, H.B., Waeyenberge, L., Spiridonov, S.E., Manandhar, H.K., and Moens, M. (2011). *Steinernema everestense* n. sp. (Rhabditida: Steinernematidae), a new species of entomopathogenic nematode from Pakhribas, Dhunkuta, Nepal. Nematology *13*, 443–462.

99. Nguyen, K.B., and Buss, E.A. (2011). *Steinernema phyllophagae* n. sp. (Rhabditida: Steinernematidae), a new entomopathogenic nematode from Florida, USA. Nematology *13*, 425–442.

100. Nguyen, K.B., Hunt, D.J., and Mracek, Z. (2007). Steinernematidae: species and descriptions. In Entomopathogenic Nematodes: Systematics, Phylogeny and Bacterial Symbionts, K.B. Nguyen and D.J. Hunt, eds. (Boston: Brill), pp. 121–609.

101. Nguyen, K.B., Puza, V., and Mracek, Z. (2008). *Steinernema cholashanense* n. sp. (Rhabditida, Steinernematidae) a new species of entomopathogenic nematode form the province of Sichuan, Chola Shan Mountains, China. J. Invertebr. Pathol. *97*, 251–264.

102. Nguyen, K.B., Stuart, R.J., Andalo, V., Gozel, U., and Rogers, M.E. (2007). *Steinernema texanum* n. sp. (Rhabditida: Steinernematidae), a new entomopathogenic nematode from Texas, USA. Nematology *9*, 379–396.

103. Spiridonov, S.E., Waeyenberge, L., and Moens, M. (2010). *Steinernema schliemanni* sp. n. (Steinernematidae; Rhabditida): a new species of steinernematids of the 'monticolum' group from Europe. Russian Journal of Nematology *18*, 175–190.

104. Stokwe, N.F., Malan, A.P., Nguyen, K.B., Knoetze, R., and Tiedt, L. (2011). *Steinernema citrae* n. sp. (Rhabditida: Steinernematidae), a new entomopathogenic nematode from South Africa. Nematology *13*, 569–587.

105. Tarasco, E., Mracek, Z., Nguyen, K.B., and Triggiani, O. (2008). *Steinernema ichnusae* sp. n. (Nematoda: Steinernematidae) a new entomopathogenic nematode from Sardinia Island (Italy). J. Invertebr. Pathol. *99*, 173–185.

106. Poinar, G.O., Jr. (1979). Nematodes for biological control of insects (Boca Raton: CRC Press).

107. Frank, J.H. (2009). Steinernema scapterisci as a biological control agent of Scapteriscus mole crickets. In Use of microbes for control and eradication of invasive arthropods, Volume 6, A.E. Hajek, ed. (The Netherlands: Springer), pp. 115–131.

108. Nguyen, K.B., and Smart, G.C. (1991). Pathogenicity of Steinernema scapterisci to Selected Invertebrates. J. Nematol. *23*, 7–11.

109. Toubarro, D., Lucena-Robles, M., Nascimento, G., Santos, R., Montiel, R., Verissimo, P., Pires, E., Faro, C., Coelho, A.V., and Simoes, N. (2010). Serine Protease-mediated Host Invasion by the Parasitic Nematode Steinernema carpocapsae. J Biol Chem *285*, 30666–30675.