Worm 1:1, 15–21; January/February/March 2012; © 2012 Landes Bioscience

WormBase
Annotating many nematode genomes

Kevin Howe,[1,*] Paul Davis,[1] Michael Paulini,[1] Mary Ann Tuli,[1] Gary Williams,[1] Karen Yook,[2] Richard Durbin,[3] Paul Kersey[1] and Paul W. Sternberg[2,*]

[1]European Bioinformatics Institute; Wellcome Trust Genome Campus; Hinxton, Cambridge UK; [2]California Institute of Technology; Division of Biology; Pasadena, CA USA; [3]Wellcome Trust Sanger Institute; Wellcome Trust Genome Campus; Hinxton, Cambridge UK

Keywords: *Caenorhabditis elegans*, nematode, genome, annotation, model organism database, community resource, sequence curation, parasitic nematode

Abbreviations: ModENCODE, Model Organism Database ENCyclopedia Of DNA Elements; EST, Expressed Sequence Tag; cDNA, complementary DNA; RNASeq, RNA sequencing by 2nd generation technologies; C., Caenorhabditis; INSDC, International Nucleotide Sequence Database Collaboration

WormBase (www.wormbase.org) has been serving the scientific community for over 11 years as the central repository for genomic and genetic information for the soil nematode *Caenorhabditis elegans*. The resource has evolved from its beginnings as a database housing the genomic sequence and genetic and physical maps of a single species, and now represents the breadth and diversity of nematode research, currently serving genome sequence and annotation for around 20 nematodes. In this article, we focus on WormBase's role of genome sequence annotation, describing how we annotate and integrate data from a growing collection of nematode species and strains. We also review our approaches to sequence curation, and discuss the impact on annotation quality of large functional genomics projects such as modENCODE.

Introduction

WormBase seeks to present an integrative view of nematode biology by in-depth curation of the research on *C. elegans* and other members of this animal family. To this end we integrate genomic sequences and annotations with curated data from genetic, developmental, physiological, behavioral and evolutionary studies. We provide multiple streams of access to the data, including the main website portal (www.wormbase.org), genome browsers, sequence search services, and application programming interfaces. WormBase aims to be the central repository and portal for nematode genomic data.

The activities of the WormBase consortium can be broadly classified into three groups: (1) curation of *C. elegans* literature and associated research and development; (2) user interface design, development and maintenance and (3) genome sequence annotation, analysis and comparative genomics. The volume of nematode data has exploded in recent years, and WormBase has had to respond accordingly in all three of these areas.[1,2] For example, as the volume and variety of information has increased, its presentation to the community in a clear and accessible way requires new forms of display. We have responded to this challenge by completely redesigning the WormBase web-interface (Harris et al., manuscript in preparation). In this article, we focus on our remit to provide integrated, coherent genome annotation for a large (and growing) collection of nematode genome sequences and strains. We also summarize our release production cycle and analysis pipelines, and describe how they affect the timeline between data submission and its subsequent public release.

Integrating and Annotating Multiple Nematode Genomes

WormBase now hosts genomic data for nearly 20 nematodes (see **Table 1**, and refs. 3–14), representing species of evolutionary, biomedical and agricultural interest. Recent additions include the parasitic nematodes *Trichinella spiralis*,[3] *Ascaris suum*[4] and *Bursaphelenchus xylophilus*.[5] The maturity of genome sequence and annotation in WormBase varies widely between species. At one end of the spectrum is the *C. elegans* genome, which was completed over a number of years using traditional physical mapping and clone-by-clone sequencing and finishing,[6] and which has highly curated annotation. More recently we have seen a number of genome sequences generated by new high-throughput low-cost technologies and many of these genomes are inevitably fragmented and incomplete; additionally, there is relatively little published functional information about many of these species.

*Correspondence to: Paul W. Sternberg and Kevin Howe; Email: pws@caltech.edu and kevin.howe@wormbase.org
Submitted: 11/02/11; Revised: 02/02/12; Accepted: 02/02/12
http://dx.doi.org/10.4161/worm.19574

Table 1. Nematode genomes in WormBase

Species	Bclade[a]	Mode of reproduction[b]	Reference strain sequenced	Integrated into WormBase	WS230 status					
					Assembly version[e]	Sequenced genome size (Mb)	Top-level fragments	Scaffold N50[f]	CDS models (distinct loci)	Gene set status
C. elegans	V	androdioecious	Bristol N2	WS1	WS230	100.3	6	17493793	25634 (20517)	Curated
C. briggsae	V	androdioecious	AF16	WS132	CAAC03000000	108.4	367	17485439	21961 (21936)	Curated
C. remanei	V	gonochoristic	PB4641	WS185	AAGD02000000	145.5	3670	461060	31476 (31471)	Curated
Brugia malayi	III	gonochoristic	TRS	WS185	AAQA01000000	95.8	27210	37841	21332 (18348)	External[g]
Pristionchus pacificus	V	hermaphoditic	PS312	WS194	v2 (Sep. 2010)	172.5	18083	1244534	24217 (24216)	External
C. japonica	V	gonochoristic	DF5080	WS195	ABLE03000000	166.3	18817	94149	36105 (29962)	Curated
C. brenneri	V	gonochoristic	PB2801	WS196	ABEG02000000	190.4	3305	368319	30670 (30667)	Curated
Meloidogyne hapla	IV	gonochoristic	VW9	WS204	ABLG01000000	53.0	3452	84000	13072 (13072)	External
Meloidogyne incognita	IV	gonochoristic	Morelos	WS205	CABB01000000	82.1	9538	83000	-	External[h]
Hemonchus contortus	V	gonochoristic	MHco3 (ISE)	WS208	v1 (Aug. 2008)	298.0	59707	13338	6201 (6201)	WormBase Predicted
C. angaria	V	gonochoristic	PS1010	WS218	AEHI01000000[i]	79.8	33559	9453	26265 (22622)	External
Trichinella spiralis	I	gonochoristic	ISS 195	WS225	ABIR02000000	63.5	6863	6373445	16380 (16380)	External
C. sp 9	V	gonochoristic	JU1422	WS226	v1 (June 2011)	204.3	7636	196652	45167 (45167)	WormBase predicted
C. sp 11	V	androdioecious	JU1373	WS226	AEKS01000000	79.3	665	20921866	27721 (22326)	External
Strongyloides ratti	IV	gonochoristic and parthenogenetic[c]	ED321 Heterogonic	WS226	CACX01000000	52.6	2184	359029	8188 (8077)	WormBase Predicted
Ascaris suum	III	gonochoristic	Natural isolate	WS229	v1 (Aug. 2011)	272.8	29831	407899	18449 (18449)	External
Bursaphelenchus xylophilus	IV	gonochoristic	Ka4C1	WS229	CADV01000000	74.6	5527	1158000	18074 (18074)	External
Heterorhabditis bacteriophora	V	gonochoristic and hermaphroditic[c,d]	M31e	WS229	ACKM01000000	77.0	1240	312328	-	External[h]
C. sp 5	V	gonochoristic	DRD-2008 JU800	WS230	v1 (Jan. 2012)	131.8	15261	25228	46280 (34696)	External

Notes: (a) ref. 15; (b) refs. 16–25; (c) heterogonic; (d) sex also determined by the environment; (e) INSDC assembly accession where available; (f) http://www.broadinstitute.org/crd/wiki/index.php/N50; (g) author gene-set extended by additional isoform predictions from WormBase; (h) awaiting submission of gene set; (i) an improved *C. angaria* assembly will be available in WS231.

WormBase undertakes different responsibilities for each of these species, which can include (1) administration of the genome sequence; (2) curation of gene models and other sequence features; (3) curation of non-sequence-based data from the literature and (4) tracking of identifiers forward through different versions of the genome sequence and annotation. The specific way in which we manage the data for a species depends (primarily) on whether we curate gene models and other features for it. It is therefore useful for the sake of discussion to classify the species into two groups: core (WormBase curated gene models) and non-core. As of release WS230, the core species are *C. elegans, C. briggsae, C. remanei, C.brenneri* and *C. japonica.*

Analyzing and presenting data for an ever-increasing number of nematode genomes requires methods that scale well. We deploy a standard automatic analysis pipeline to annotate all the species we house (core and non-core), including repeat prediction, cDNA alignments, the determination of homology relationships, and protein domain identification. If a genome sequence for a non-core species is submitted without a gene-set, we also run an in-house gene prediction pipeline that uses CEGMA[26] to accurately identify a small, universally conserved set of gene models. These are then used to train parameters for AUGUSTUS,[27] which we then apply using protein homologies and any available RNASeq and other transcript data as supporting evidence. In some cases, these internally-produced gene predictions are later replaced by a canonical set of models provided by the submitters.

Updating an existing species in WormBase with a new assembly and/or gene-set presents additional challenges, because users rely on stable identifiers to track their entities of interest, which must be propagated forward to corresponding features in subsequent releases. For core species, identifiers are actively managed and tracked using our own curation software infrastructure. For non-core species, we use the Ensembl[28] stable-identifier mapping software for this task.

The principal way in which we draw information from multiple species together is by connecting genes via orthology and paralogy relationships to genes in other species (both nematode and other model organisms such as human, mouse and fly). As of WS230, we include relationships published by the following projects and resources: InParanoid[29] (version 7); TreeFam[30] (version 7); the Othologous Matrix Project[31] (OMA, August 2009/08 version); OrthoMCL;[32] PantherDB[33,34] (version 7); and Ensembl[28,35] (version 65). In addition, we curate orthology calls from the literature (e.g., Hillier et al., ref. 8) and direct submissions. We also use data in eggNOG[36] (version 3.0) to cluster genes into functionally characterized homologous groups.

These resources are inevitably based on snapshots of the gene models, taken at various times. For our core species however, particularly *C. elegans*, the gene models are in a state of flux, being revised and improved on the basis of the latest evidence. In order to infer up-to-date nematode homology relationships for the latest gene models, we run the Ensembl Compara GeneTree pipeline[35] as part of the preparation for every WormBase release. The resulting gene trees are used to infer additional current orthology relationships to those obtained by import from the third-party resources and direct submission.

One way in which we use the orthology relationships internally is to project WormBase-approved gene names[37] onto orthologous gene(s) of other nematode species. For this a conservative approach is adopted: each proposed gene name is required to be supported by an unambiguous one to one orthology connection according to the majority of available source analyses.

We also use Ensembl Compara DNA pipeline[38] to produce whole-genome multiple alignments of all genomes in WormBase and derived genome conservation tracks (using GERP[39]). However, as the genetic diversity of the species collection in WormBase continues to increase, a single multiple alignment for all nematodes becomes less appropriate. We therefore propose to replace it with a series of pairwise alignments, providing multiple alignments only for selected subsets of species.

Sequence Curation

WormBase adopts an anomaly-driven approach to curation, whereby discrepancies between current gene models and alignment data are identified and flagged as curation targets. We have implemented a software application (CurationTool) that identifies these discrepancies and scores them according to their degree of discordance, presenting the results to the curator using a graphical user interface. An in-depth discussion of CurationTool and our anomaly-driven curation is presented elsewhere.[40]

For protein-coding genes, WormBase curates only the protein-coding portion (CDS) of the full transcript. For our core species, we use the high-confidence subset of cDNA alignments overlaying the curated CDS models to infer a set of full-length transcripts (including 5' and 3' untranslated regions), using a custom algorithm (unpublished). In the past, the accuracy of this process has been sensitive to artifacts such as alignment errors or chimeric cDNAs, but we have recently improved the algorithm to take these factors into account.

The primary line of evidence for gene model curation is transcript data. In addition to cDNAs deposited in the nucleotide archives, we draw data from numerous resources, publications and direct submissions. We also align all RNASeq data deposited in the Short Read Archive (SRA) to our core species using TopHat,[41] and infer gene expression estimates for a variety of life stages and environmental conditions using Cufflinks.[42]

WormBase is committed to act as the ultimate repository for data coming from the nematode half of the modENCODE[43,44] project. Most data sets have been accessible via the genome browser since the summer of 2010. To extract the maximum utility from the data, it is integrated fully into our database, by extending the data models where necessary and adding full cross-referencing and connectivity with existing WormBase objects. To date, the focus for full integration has been on data sets with high impact on gene model and other sequence feature curation, namely: trans-splice sites;[45] poly-A cleavage sites and untranslated regions;[44,46] large-scale EST sets (P. Green; data retrieved from nucleotide archives); mass-spectrometry peptide sequences;[44] and RNASeq transcripts, and derived gene-predictions.[44]

The data of highest impact for curation has been the RNASeq transcriptome, and this has been used in a number of different

ways. First, the modENCODE "genelets" (fragmentary gene models constructed using RNASeq data from 14 life stages) have been used to produce a new anomaly type for CurationTool that highlights potential cases where adjacent genes could be merged. To date, over three hundred cases displaying this anomaly have been scrutinized, of which approximately 35% resulted in a merge, and a further 10% some other change (for example the movement of an exon from one gene to another). Second, we have re-visited the source RNASeq data and analyzed it using the Tophat/Cufflnks pipeline[41,42] to identify candidate "RNASeq-splice" features. These can be used both to confirm introns already part of curated gene models, and also to suggest changes to existing gene models or new isoforms. Third, the strand bias characteristic of the modENCODE RNASeq alignments[47] has been extremely useful for curators to resolve ambiguities in the definition of the 5' and 3' ends of genes. Finally, the modENCODE RNASeq data has allowed us to make corrections to the *C. elegans* reference genome itself. By taking proposed errors and verifying them using data from a private submission of high-throughput-sequencing (J. Ahringer and M. Berriman, pers. comm.), we have been able to make 156 genome sequence corrections (110 insertions, 44 deletions and 2 substitutions), resulting in the correction of 100 gene models.

Additionally, since the data from modENCODE began to become available from the project Data Co-ordination Centre, the following data sets have been subjected to rigorous internal quality control and fully integrated into the database: ~300 Highly Occupied Target (HOT) regions;[44] ~7,000 non-coding RNA genes;[44] the probable parent for ~1,000 pseudogenes;[44] and ~21,000 three-prime UTRs from the UTRome project.[46] We will prioritise the incorporation of the transcription-factor binding site and chromatin accessibility data as soon as the final versions of these data sets are made available.

We have also worked with groups performing their own analysis of the modENCODE data. For example, a study of the modENCODE RNASeq reads (T. Blumenthal, pers. comm.) has resulted in significant improvements to the operon data set. This has involved identifying cases where fewer than 5% of the trans-splice leader reads for "internal" genes (i.e., genes other than the first) were SL2 type, and modifying the gene content of the operons accordingly.

In addition to modENCODE, we continue to draw in data from the scientific literature and direct submissions, often combining different data sources to assist in making correct predictions. The modENCODE poly-A site data has been supplemented with a corresponding data set from an independent study.[48] These two data sets have only 25% redundancy, and over 80% of coding genes now have an annotated polyA site in WormBase. Gene predictions by genBlastG[49] based on BLAST homologies to *C. elegans* proteins have also proved valuable for the curation of *C. briggsae, C. brenneri,* and *C. remenei*.

We can assess gene-model accuracy in the presence of fragmentary transcript evidence by measuring the proportion of curated introns that are confirmed by spliced cDNA evidence. For WS230, the proportion of *C. elegans* curated CDS introns confirmed by traditional cDNA, modENCODE RNASeq and mass-spectrometry evidences is 83%, 88% and 14% respectively. Overall, 93% of curated introns are confirmed and 82% of CDS models have all of their introns confirmed by at least one of these three lines of evidence; the corresponding measurements for the final release prior to modENCODE (WS200, February 2009) were 74% and 56%, demonstrating the value of the project in increasing the accuracy and confidence of *C. elegans* gene models.

Intraspecies Variation

Similar to many other resources, WormBase captures within-species variation as differences (insertions, deletions and substitutions) with respect to the genome sequence of the reference strain. We expect variation data for many nematode species in the future, but at present almost all the data we house is for *C. elegans*.

Historically, the majority of variation data we have processed has been from laboratory-manipulated strains. We maintain close working relationships and established data exchange protocols with the *Caenorhabditis* Genetics Center (CGC; www.cbs.umn.edu/CGC), the *C. elegans* Gene Knockout Consortium (GKC; www.celeganskoconsortium.omrf.org), and the National BioResource Project of Japan (NBRP; www.shigen.nig.ac.jp/c.elegans/index.jsp). We also curate variation data from individual user submissions; which although time-consuming, are often biologically important.

There has recently been a rapid growth of *C. elegans* variation data generated by whole genome sequencing projects (refs. 50–54; Andersen et al., manuscript in preparation; Moerman and Waterston, manuscript in preparation). These data sets include an increasing number of variations from naturally-occurring wild-isolate strains. Motivated by community feedback, we have increased the clarity of our representation and display of this information. Every variation object processed by WormBase is assigned a unique, stable identifier with prefix "WBVar." For laboratory-induced variations, we also assign a more directly informative public name comprised of a project/laboratory prefix (supplied by J. Hodgkin, pers. comm.) and a numerical suffix. For naturally occurring variations, the public name defaults to the WBVar identifier, making the distinction between these objects and the laboratory induced variations obvious and immediate.

We now also collect non-sequence-based information for wild isolate strains (http://tazendra.caltech.edu/~azurebrd/cgi-bin/forms/wild_isolate.cgi). Compared with laboratory-manipulated strains, there is additional information to capture about the wild isolates, such as isolation location, the condition in which it was found, and details of how it was isolated. Many wild isolates are not stocked at the CGC, and WormBase acts as the central data repository for these strains.

WormBase does not have a mandate to act as a permanent repository for variation data, and as the volume of these data sets continues to rapidly increase, we become less adequately resourced to perform this function. Projects are therefore encouraged to submit their data to the NCBI's Database of Short Genetic

Variations (dbSNP),[55] an established archive for variation data. We act as a submission broker in cases where a laboratory lacks the technical resources to conform to the dbSNP submission protocols. To date, data from six projects have been integrated into WormBase and submitted to dbSNP. WormBase adds value to these data sets by performing additional analysis and placing them into context with other data types (e.g., Gene).

Variations are most often submitted to WormBase as a molecular change at given location in a specific version of the reference genome sequence. As part of the curation, we capture and record a short flanking sequence either side of the variation feature, disassociating it from a specific version of the reference genome. Each release, we re-map all variations and re-calculate potential consequences of the molecular changes (e.g., non-sense, mis-sense or silent protein-coding mutation) on the latest gene models.

Release Cycle and Database Build

WormBase is released every two months, with the preparation for a release beginning three months in advance. This release cycle can give rise to variability in the time between a curator transaction (e.g., the update of a gene name, correction of an error, or the import of a new data set) and its availability on the WormBase website. The delay can be as short as three months (if the change is made immediately before we start building the release) and as long as five months (if made immediately after, in which case it will not be public until the following release).

Building a WormBase database release is a complicated process, the broad stages of which can be described as: (1) data freeze, where each contributing consortium partner takes a snap-shot of the database(s) in which their curation data are stored; (2) data collation, where the curation database snap-shots are brought together into a single database; (3) submission of updated annotation on core species to the International Nucleotide Sequence Database Collaboration,[56] to ensure that the representation of core nematode data in the nucleotide and protein archives is up-to-date; (4) mapping of sequence data (e.g., cDNAs, microarray probes, sequence features, variations) to the genome; (5) establishing connections between objects of different types (e.g., RNAi to Gene), usually via genomic location; (6) the large-scale computational analyses discussed earlier, such as homology detection and whole-genome alignment; and (7) quality control and assurance.

For the more complicated parts of the build process, we deploy two components of the Ensembl system for the management and tracking of computational pipelines: ensembl-pipeline[57] for homology analysis and eHive[58] for comparative analysis. The key features of these systems are (1) automatic re-run of tasks that have failed; and (2) user-definition of a sub-task dependency graph for a process, allowing complex pipelines to be run with minimal user intervention. These systems are critical in enabling us to produce the database in a regular and timely manner.

Each stage of the database production is subject to a suite of integrity checks to ensure that it has completed cleanly and without error. For example, we compare the number of objects in each data class with the count at the corresponding stage in the previous release. Major discrepancies are flagged for investigation. This mechanism has proved to be extremely effective in catching errors and process failures as soon as they occur.

Summary

WormBase is facing a deluge of data from many nematode genome sequencing projects, and we have prepared for this by putting into place annotation and integration pipelines and workflows that will allow the data to be analyzed and presented in a timely and consistent manner. As ever, we welcome feedback and ideas from our user-base as part of the continued development of the resource. We are currently particularly interested in suggestions on how we can maximise the utility of housing a broad representation of the nematode phylum, and what comparative genomics services and views users would find most useful. Users can contact the developers at help@wormbase.org with their suggestions.

Acknowledgments

This work is supported by the US National Institutes of Health (Grant no. P41 HG02223); US National Human Genome Research Institute (Grant no. P41-HG02223); and British Medical Research Council (Grant no. G070119); P.W.S. is an investigator with the Howard Hughes Medical Institute. Funding for open access charge: US National Human Genome Research Institute (Grant no. P41-HG02223).

References

1. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, et al. WormBase 2012: more genomes, more data, new website. Nucleic Acids Res 2012; 40(Database issue):D735-41; PMID:22067452; http://dx.doi.org/10.1093/nar/gkr954
2. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res 2010; 38(Database issue):D463-7; PMID:19910365; http://dx.doi.org/10.1093/nar/gkp952
3. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, et al. The draft genome of the parasitic nematode *Trichinella spiralis*. Nat Genet 2011; 43:228-35; PMID:21336279; http://dx.doi.org/10.1038/ng.769
4. Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, et al. *Ascaris suum* draft genome. Nature 2011; 479:529-33; PMID: 22031327; http://dx.doi.org/10.1038/nature10553
5. Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, et al. Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. PLoS Pathog 2011; 7: e1002219; PMID:21909270; http://dx.doi.org/10.1371/journal.ppat.1002219
6. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 1998; 282:2012-8; PMID:9851916; http://dx.doi.org/10.1126/science.282.5396.2012
7. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, et al. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol 2003; 1:E45; PMID:14624247; http://dx.doi.org/10.1371/journal.pbio.0000045
8. Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, et al. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol 2007; 5:e167; PMID:17608563; http://dx.doi.org/10.1371/journal.pbio.0050167

9. Ross JA, Koboldt DC, Staisch JE, Chamberlin HM, Gupta BP, Miller RD, et al. *Caenorhabditis briggsae* recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination. PLoS Genet 2011; 7:e1002174; PMID:21779179; http://dx.doi.org/10.1371/journal.pgen.1002174
10. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, et al. Draft genome of the filarial nematode parasite *Brugia malayi*. Science 2007; 317:1756-60; PMID:17885136; http://dx.doi.org/10.1126/science.1145406
11. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, et al. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat Genet 2008; 40:1193-8; PMID:18806794; http://dx.doi.org/10.1038/ng.227
12. Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EG, Deleury E, et al. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nat Biotechnol 2008; 26:909-15; PMID:18660804; http://dx.doi.org/10.1038/nbt.1482
13. Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, et al. Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. Proc Natl Acad Sci U S A 2008; 105:14802-7; PMID:18809916; http://dx.doi.org/10.1073/pnas.0805946105
14. Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, et al. Scaffolding a Caenorhabditis nematode genome with RNA-seq. Genome Res 2010; 20:1740-7; PMID:20980554; http://dx.doi.org/10.1101/gr.111021.110
15. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, et al. A molecular evolutionary framework for the phylum Nematoda. Nature 1998; 392:71-5; PMID:9510248; http://dx.doi.org/10.1038/32160
16. Haag S. The evolution of nematode sex determination: *C. elegans* as a reference point for comparative biology (December 29 2005). In: The *C. elegans* Research Community ed. WormBook. http://www.wormbook.org
17. Kiontke KC, Félix MA, Ailion M, Rockman MV, Braendle C, Pénigault JB, et al. A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. BMC Evol Biol 2011; 11:339; PMID:22103856; http://dx.doi.org/10.1186/1471-2148-11-339
18. Mayer WE, Herrmann M, Sommer RJ. Phylogeny of the nematode genus Pristionchus and implications for biodiversity, biogeography and the evolution of hermaphroditism. BMC Evol Biol 2007; 7:104; PMID:17605767; http://dx.doi.org/10.1186/1471-2148-7-104
19. Redman E, Grillo V, Saunders G, Packard E, Jackson F, Berriman M, et al. Genetics of mating and sex determination in the parasitic nematode *Haemonchus contortus*. Genetics 2008; 180:1877-87; PMID:18854587; http://dx.doi.org/10.1534/genetics.108.094623
20. Bird DM, Williamson VM, Abad P, McCarter J, Danchin EG, Castagnone-Sereno P, et al. The genomes of root-knot nematodes. Annu Rev Phytopathol 2009; 47:333-51; PMID:19400640; http://dx.doi.org/10.1146/annurev-phyto-080508-081839
21. Ciche T. The biology and genome of *Heterorhabditis bacteriophora* (February 20 2007). In: The *C. elegans* Research Community ed. WormBook. http://www.wormbook.org
22. Viney ME. A genetic analysis of reproduction in *Strongyloides ratti*. Parasitology 1994; 109:511-5; PMID:7800419; http://dx.doi.org/10.1017/S0031182000080768
23. Pires-daSilva A. Evolution of the control of sexual identity in nematodes. Semin Cell Dev Biol 2007; 18:362-70; PMID:17306573; http://dx.doi.org/10.1016/j.semcdb.2006.11.014
24. Boag PR, Newton SE, Gasser RB. Molecular aspects of sexual development and reproduction in nematodes and schistosomes. Adv Parasitol 2001; 50:153-98; PMID:11757331; http://dx.doi.org/10.1016/S0065-308X(01)50031-7
25. Hasegawa K, Mota MM, Futai K, Miwa J. Chromosome structure and behaviour in *Bursaphelenchus xylophilus* (Nematoda: Parasitaphelenchidae) germ cells and early embryo. Nematology 2006; 8:425-34; http://dx.doi.org/10.1163/156854106778493475
26. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007; 23:1061-7; PMID:17332020; http://dx.doi.org/10.1093/bioinformatics/btm071
27. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 2006; 7:62; PMID:16469098; http://dx.doi.org/10.1186/1471-2105-7-62
28. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, et al. Ensembl 2011. Nucleic Acids Res 2011; 39(Database issue):D800-6; PMID:21045057; http://dx.doi.org/10.1093/nar/gkq1064
29. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 2010; 38(Database issue):D196-203; PMID:19892828; http://dx.doi.org/10.1093/nar/gkp931
30. Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 2006; 34(Database issue):D572-80; PMID:16381935; http://dx.doi.org/10.1093/nar/gkj118
31. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 2011; 39(Database issue):D289-94; PMID:21113020; http://dx.doi.org/10.1093/nar/gkq1238
32. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 2006; 34(Database issue):D363-8; PMID:16381887; http://dx.doi.org/10.1093/nar/gkj123
33. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res 2003; 31:334-41; PMID:12520017; http://dx.doi.org/10.1093/nar/gkg115
34. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 2010; 38(Database issue):D204-10; PMID:20015972; http://dx.doi.org/10.1093/nar/gkp1019
35. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 2009; 19:327-35; PMID:19029536; http://dx.doi.org/10.1101/gr.073585.107
36. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, et al. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 2010; 38(Database issue):D190-5; PMID:19900971; http://dx.doi.org/10.1093/nar/gkp951
37. Horvitz HR, Brenner S, Hodgkin J, Herman RK. A uniform genetic nomenclature for the nematode *Caenorhabditis elegans*. Mol Gen Genet 1979; 175:129-33; PMID:292825; http://dx.doi.org/10.1007/BF00425528
38. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res 2008; 18:1814-28; PMID:18849524; http://dx.doi.org/10.1101/gr.076554.108
39. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A, et al.; NISC Comparative Sequencing Program. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005; 15:901-13; PMID:15965027; http://dx.doi.org/10.1101/gr.3577405
40. Williams GW, Davis PA, Rogers AS, Bieri T, Ozersky P, Spieth J. Methods and strategies for gene structure curation in WormBase. Database (Oxford) 2011; 2011:baq039; PMID:21543339; http://dx.doi.org/10.1093/database/baq039
41. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25:1105-11; PMID:19289445; http://dx.doi.org/10.1093/bioinformatics/btp120
42. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010; 28:511-5; PMID:20436464; http://dx.doi.org/10.1038/nbt.1621
43. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, et al.; modENCODE Consortium. Unlocking the secrets of the genome. Nature 2009; 459:927-30; PMID:19536255
44. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al.; modENCODE Consortium. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 2010; 330:1775-87; PMID:21177976; http://dx.doi.org/10.1126/science.1196914
45. Allen MA, Hillier LW, Waterston RH, Blumenthal T. A global analysis of *C. elegans trans*-splicing. Genome Res 2011; 21:255-64; PMID:21177958; http://dx.doi.org/10.1101/gr.113811.110
46. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, et al. The landscape of C. elegans 3'UTRs. Science 2010; 329:432-5; PMID:20522740; http://dx.doi.org/10.1126/science.1191244
47. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. Genome Res 2009; 19:657-66; PMID:19181841; http://dx.doi.org/10.1101/gr.088112.108
48. Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature 2011; 469:97-101; PMID:21085120; http://dx.doi.org/10.1038/nature09616
49. She R, Chu JS, Uyar B, Wang J, Wang K, Chen N. genBlastG: using BLAST searches to build homologous gene models. Bioinformatics 2011; 27:2141-3; PMID:21653517; http://dx.doi.org/10.1093/bioinformatics/btr342
50. Zuryn S, Le Gras S, Jamet K, Jarriault S. A strategy for direct mapping and identification of mutations by whole-genome sequencing. Genetics 2010; 186:427-30; PMID:20610404; http://dx.doi.org/10.1534/genetics.110.119230
51. Sarin S, Bertrand V, Bigelow H, Boyanov A, Doitsidou M, Poole RJ, et al. Analysis of multiple ethyl methanesulfonate-mutagenized *Caenorhabditis elegans* strains by whole-genome sequencing. Genetics 2010; 185:417-30; PMID:20439776; http://dx.doi.org/10.1534/genetics.110.116319
52. Flibotte S, Edgley ML, Chaudhry I, Taylor J, Neil SE, Rogula A, et al. Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. Genetics 2010; 185:431-41; PMID:20439774; http://dx.doi.org/10.1534/genetics.110.116616
53. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. Nat Methods 2008; 5:865-7; PMID:18677319; http://dx.doi.org/10.1038/nmeth.1249

54. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods 2008; 5:183-8; PMID:18204455; http://dx.doi.org/10.1038/nmeth.1179
55. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001; 29:308-11; PMID:11125122; http://dx.doi.org/10.1093/nar/29.1.308
56. Karsch-Mizrachi I, Nakamura Y, Cochrane G; International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2012; 40(Database issue):D33-7; PMID:22080546; http://dx.doi.org/10.1093/nar/gkr1006
57. Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, et al. The Ensembl analysis pipeline. Genome Res 2004; 14:934-41; PMID:15123589; http://dx.doi.org/10.1101/gr.1859804
58. Severin J, Beal K, Vilella AJ, Fitzgerald S, Schuster M, Gordon L, et al. eHive: an artificial intelligence workflow system for genomic analysis. BMC Bioinformatics 2010; 11:240; PMID:20459813; http://dx.doi.org/10.1186/1471-2105-11-240