25. T. M. Barnes, Y. Jin, H. R. Horvitz, G. Ruvkun, S. Hekimi, *J. Neurochem.* **67**, 46 (1996).
26. C. M. Coburn and C. I. Bargmann, *Neuron* **17**, 695 (1996).
27. H. Komatsu, I. Mori, J.-S. Rhee, N. Akaike, Y. Ohshima, *ibid.*, p. 707.
28. M. J. Caterina *et al.*, *Nature* **389**, 816 (1997).
29. H. A. Colbert, T. L. Smith, C. I. Bargmann, *J. Neurosci.* **17**, 8259 (1997).
30. A. O. W. Stretton, R. E. Davis, J. D. Angstadt, J. E. Donmoyer, C. D. Johnson, *Trends Neurosci.* **8**, 294 (1985).
31. J. P. Walrond, I. S. Kass, A. O. W. Stretton, J. E. Donmoyer, *J. Neurosci.* **5**, 1 (1985).
32. L. Avery and H. R. Horvitz, *J. Exp. Zool.* **253**, 263 (1990).
33. D. M. Raizen, R. Y. N. Lee, L. Avery, *Genetics* **141**, 1365 (1995).
34. H. Li, L. Avery, W. Denk, G. P. Hess, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5912 (1997).
35. A. Alfonso, K. Grundahl, J. R. McManus, J. B. Rand, *J. Neurosci.* **14**, 2290 (1994).
36. A. Alfonso, K. Grundahl, J. S. Duerr, H. P. Han, J. B. Rand, *Science* **261**, 617 (1993).
37. C. D. Johnson *et al.*, *Genetics* **97**, 261 (1981).
38. M. Arpagaus *et al.*, *J. Biol. Chem.* **269**, 9957 (1994).
39. S. L. McIntire, R. J. Reimer, K. Schuske, R. H. Edwards, E. M. Jorgensen, *Nature* **389**, 870 (1997).
40. A. O. Stretton, C. Cowden, P. Sithigorngul, R. E. Davis, *Parasitology* **102**, S107 (1991).
41. C. Cowden, P. Sithigorngul, P. Brackley, J. Guastella, A. O. Stretton, *J. Comp. Neurol.* **333**, 455 (1993).
42. L. S. Nelson, L. Kim, J. E. Memmott, C. Li, *Mol. Brain Res.* **58**, 103 (1998).
43. L. S. Nelson, M. L. Rosoff, C. Li, *Science* **281**, 1686 (1998).
44. I. N. Maruyama and S. Brenner, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 5729 (1991).
45. K. Gengyo-Ando *et al.*, *Neuron* **11**, 703 (1993).
46. E. M. Jorgensen and M. L. Nonet, *Semin. Dev. Biol.* **6**, 207 (1995).
47. J. B. Rand and C. D. Johnson, in *Methods in Cell Biology*. Caenorhabditis elegans: *Modern Biological Analysis of an Organism*, H. F. Epstein and D. C. Shakes, Eds. (Academic Press, San Diego, 1995), vol. 48, pp. 187–204.
48. K. Iwasaki, J. Staunton, O. Saifee, M. Nonet, J. H. Thomas, *Neuron* **18**, 613 (1997).
49. M. L. Nonet, K. Grundahl, B. J. Meyer, J. B. Rand, *Cell* **73**, 1291 (1993).
50. E. M. Jorgensen *et al.*, *Nature* **378**, 196 (1995).
51. M. L. Nonet *et al.*, *J. Neurosci.* **17**, 8061 (1997).
52. M. L. Nonet, O. Saifee, H. Zhao, J. B. Rand, L. Wei, *ibid.* **18**, 70 (1998).
53. O. Saifee, L. Wei, M. L. Nonet, *Mol. Biol. Cell* **9**, 1235 (1998).
54. For example, all the exocytosis proteins except synaptotagmin are conserved in yeast secretory pathways.
55. J. P. Walrond and A. O. W. Stretton, *J. Neurosci.* **5**, 9 (1985).
56. J. T. Fleming *et al.*, *ibid.* **17**, 5843 (1997).
57. M. Treinin and M. Chalfie, *Neuron* **14**, 871 (1995).
58. A. Hart, S. Sims, J. Kaplan, *Nature* **378**, 82 (1995).
59. A. V. Maricq, E. Peckol, M. Driscoll, C. I. Bargmann, *ibid.*, p. 78.
60. D. F. Cully *et al.*, *ibid.* **371**, 707 (1994).
61. J. A. Dent, M. W. Davis, L. Avery, *EMBO J.* **16**, 5867 (1997).
62. S. M. Kaech, C. W. Whitfield, S. K. Kim, *Cell* **94**, 761 (1998).
63. C. Rongo, C. W. Whitfield, A. Rodal, S. K. Kim, J. M. Kaplan, *ibid.*, p. 751.
64. B. Olde and W. R. McCombie, *J. Mol. Neurosci.* **8**, 53 (1997).
65. E. L. L. Sonnhammer and R. Durbin, *Genomics* **46**, 200 (1997).
66. M. de Bono and C. I. Bargmann, *Cell* **94**, 679 (1998).
67. P. Sengupta, J. C. Chou, C. I. Bargmann, *ibid.* **84**, 899 (1996).
68. H. M. Robertson, *Genome Res.* **8**, 449 (1998).
69. E. R. Troemel, J. H. Chou, N. D. Dwyer, H. A. Colbert, C. I. Bargmann, *Cell* **83**, 207 (1995).
70. S. Yu, L. Avery, E. Baude, D. L. Garbers, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3384 (1997).
71. D. Birnby and J. Thomas, unpublished data.
72. J. Mendel *et al.*, *Science* **267**, 1652 (1995).
73. L. Segalat, D. Elkes, J. Kaplan, *ibid.*, p. 1648.
74. L. Brundage *et al.*, *Neuron* **16**, 999 (1996).
75. H. C. Korswagen, J. H. Park, Y. Ohshima, R. H. Plasterk, *Genes Dev.* **11**, 1493 (1997).
76. R. R. Zwaal, J. E. Mendel, P. W. Sternberg, R. H. A. Plasterk, *Genetics* **145**, 715 (1997).
77. K. Roayaie, J. G. Crump, A. Sagasti, C. I. Bargmann, *Neuron* **20**, 55 (1998).
78. R. Zwaal *et al.*, *Cell* **84**, 619 (1996).
79. A. J. Berger, A. C. Hart, J. M. Kaplan, *J. Neurosci.* **18**, 2871 (1998).
80. M. Chalfie *et al.*, *ibid.* **5**, 956 (1985).
81. T. A. Starich, R. Y. N. Lee, C. Panzarella, L. Avery, J. E. Shaw, *J. Cell Biol.* **134**, 537 (1996).
82. P. Phelan *et al.*, *Nature* **391**, 181 (1998).
83. T. M. Barnes and S. Hekimi, *J. Neurochem.* **69**, 2251 (1997).
84. T. A. Starich, R. K. Herman, J. E. Shaw, *Genetics* **133**, 527 (1993).
85. E. M. Hedgecock and R. L. Russell, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 4061 (1975).
86. C. H. Rankin, C. D. Beck, C. M. Chiba, *Behav. Brain Res.* **37**, 89 (1990).
87. H. A. Colbert and C. I. Bargmann, *Learn. Mem.* **4**, 179 (1997).
88. J. Y. Wen *et al.*, *Behav. Neurosci.* **111**, 354 (1997).
89. C. H. Bailey, D. Bartsch, E. R. Kandel, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13445 (1996).
90. A. J. Silva, J. H. Kogan, P. W. Frankland, S. Kida, *Annu. Rev. Neurosci.* **21**, 127 (1998).
91. E. Perez-Reyes *et al.*, *Nature* **391**, 896 (1998).

92. I thank J. Thomas for his ideas and insights into ion channels; R. Durbin and M. Finney for assistance with sequence analysis; and E. Sonnhammer, L. Salkoff, H. Robertson, D. Garbers, T. Barnes, and S. Hekimi for their analysis of various gene families. J. Thomas, V. Maricq, W. Davis, J. Dent, E. Lundquist, M. Finney, and B. Horvitz provided valuable comments on this manuscript.

The Taxonomy of Developmental Control in *Caenorhabditis elegans*

Gary Ruvkun* and Oliver Hobert

REVIEW

The *Caenorhabditis elegans* genome sequence was surveyed for transcription factor and signaling gene families that have been shown to regulate development in a variety of species. About 10 to 25 percent of the genes in most of the gene families already have been genetically analyzed in *C. elegans*, about half of the genes detect probable orthologs in other species, and about 10 to 25 percent of the genes are, at present, unique to *C. elegans*. *Caenorhabditis elegans* is also missing genes that are found in vertebrates and other invertebrates. Thus the genome sequence reveals universals in developmental control that are the legacy of metazoan complexity before the Cambrian explosion, as well as genes that have been more recently invented or lost in particular phylogenetic lineages.

Genetic analysis of development has been a traditional focus of *C. elegans* research. Approximately 200 of the 1600 loci that have been identified by genetic analysis cause the sort of cell fate transformations and patterning defects that attract developmental geneticists, and so far about 150 genes (almost 1% of the total genes) have been studied molecularly (*1*). This set of molecularly analyzed developmental control genes, while biased toward particular intensively studied pathways, represents genes that control a fairly random sample of developmental events. More than 90% are related to genes identified by analogous molecular and genetic analyses, especially in *Drosophila* and vertebrates. Most of the genes fit into the modern developmental control canon: growth factor signaling pathways (about 30% of the genes) and transcriptional regulatory cascades (about 25% of the genes). These sequence similarities allow developmental control to be described in molecular terms. Only 10% of these genes show no detectable sequence similarity to other genes in the databases. This is in contrast to the overall genome sequence, which reveals that about 50% of *C. elegans* genes encode novel proteins. The underrepresentation of novel genes in the set of developmental control genes identified by genetics, which is not biased toward any particular molecular feature, implies that a conserved set of genes regulates metazoan development.

Most of the gene families that include the genetically identified *C. elegans* control genes are large and contain members from many species; these families can be classified into dendrograms of relatedness (*2*) (Fig. 1). For example, the tree of 355 homeobox genes classifies the relatedness of an ancient, highly ramified gene family.

The authors are in the Department of Molecular Biology, Massachusetts General Hospital, Department of Genetics, Harvard Medical School, Boston, MA 02114, USA.

*To whom correspondence should be addressed. E-mail: ruvkun@frodo.mgh.harvard.edu

Genes on many branches of the tree have been shown to perform particular developmental functions. For all these dendrograms, *C. elegans* family members are distributed in most branches. In many cases, a particular protein sequence from *C. elegans* is more closely related to one particular mammalian or other invertebrate protein sequence than to any other *C. elegans* member of the family, suggesting that these proteins are orthologous. That is, the common ancestor of nematodes and the other species carried a gene of this sequence type, and its features have been maintained by similar selection in both lineages since their divergence. In a few cases, this probable orthology indicated by sequence similarity has been confirmed by showing that the genes have similar developmental roles.

Even more dramatically, exhaustive genetic analysis of particular *C. elegans* developmental pathways has revealed examples in which each gene in a regulatory cascade detects an ortholog that also acts in the equivalent regulatory cascade in phylogenetically distant species. For example, genetic analysis of signaling in one epidermal tissue of *C. elegans* revealed a mammalian epidermal growth factor (EGF)–like signaling protein, an EGF-like receptor, a Ras-like signal transduction protein, a GRB2-like adaptor protein, a Raf/Map kinase cascade, and two different transcription factors that couple to those kinases. All of the *C. elegans* genes in the pathway are orthologous to genes in mammalian and *Drosophila* signaling pathways (*3*). Similarly, CED-9 in *C. elegans* regulates an orthologous protease regulatory cascade to control cell death, and the *C. elegans* DAF-2 insulin-like receptor regulates an orthologous metabolism regulating kinase cascade (*3*). This congruence of *C. elegans* and other animal developmental control pathways has its roots in the antiquity of each pathway in multicellular developmental control.

The phylogenetic placement of *C. elegans* is crucial to the interpretation of the similarities and differences we detail below between the developmental control gene families of *C. elegans* and other animal species. The position of *C. elegans*, and more generally nematodes, in the phylogenetic trees constructed from sequence comparisons has recently been reevaluated (*4*). Consideration of mutation rate effects on phylogenetic placement by parsimony have moved the nematode branch to a cluster of molting invertebrates termed the Ecdysozoa, from its previous position branching before the chordate/articulata divergence. Thus the Nematoda are expected to have more in common with molting arthropods than chordates or the other major animal clade, the lophotrochozoans (such as segmented worms) (*4*). This phylogenetic placement predicts that developmental control genes common to other Ecdysozoa (such as *Drosophila*) and chordates (such as humans) should also be found in *C. elegans*.

Because the genetic analyses of a large variety of developmental events in vertebrate and invertebrate animals have converged on similar families of growth factor signaling and transcription cascade molecules, an exploration of the extent of each such family in *C. elegans* is a promising avenue for inferring some of what remains to be discovered. Perhaps in this way, we can infer how many mechanistically homologous but not orthologous pathways remain to be explored.

Transcriptional Regulators of Development

Consistent with a major tenet of developmental biology that distinct cell types transcribe specialized sets of genes, transcription factors have loomed large in the control genes identified by *C. elegans* (as well as *Drosophila*) developmental genetics.

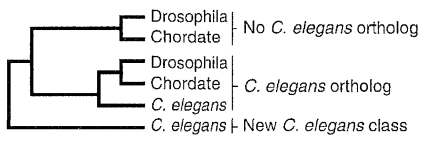

Fig. 1. Schematic dendrogram explaining how orthology statements are derived. The absence of *C. elegans* orthologs of *Drosophila*/chordate genes are significant because the complete *C. elegans* genome can be searched; however, until other metazoan genomes are sequenced, the assignment of a "new" *C. elegans* class is preliminary.

Homeobox genes. The homeobox gene family mediates a variety of developmental events across phylogeny (*5*) and has as its hallmark a particular 60 amino acid DNA-binding motif. The wide variety of homeobox genes identified over the past 15 years are expressed in distinct patterns in space, time, or cell type. They are thought to regulate patterning and cell type specification events during development by binding to distinct arrays of downstream genes to coordinate their expression. The *C. elegans* genome sequence reveals 83 homeodomain family members of many subclasses, a number near the 60 predicted from a degenerate oligonucleotide screen early in the genome project (*6*). These homeobox genes have been a major target of the *C. elegans* developmental genetic analysis to date: 17 of the initial 150 developmental genetic loci that have been studied molecularly are homeobox genes, and 7 additional homeobox genes have been analyzed by gene disruption. Thus, homeobox genes constitute about 1/10 of the *C. elegans* developmental control genes revealed so far by developmental genetics, and about 25% of the gene family has already been genetically analyzed. These genes mediate developmental processes ranging from spatial patterning by the *Hox* cluster subclass (*3*) to neural differentiation and neurotransmitter specification by the *unc-30* subclass (*7*).

There are chordate or other invertebrate orthologs of 51 of the 83 *C. elegans* homeobox genes, many of which have been genetically studied in those other systems. In some cases the orthology indicated by sequence comparison has been endorsed by genetic analysis in multiple species. For example, both the *C. elegans* homeobox gene *unc-86* and the mammalian ortholog *Brn-3* mediate the maturation of mechanosensory neurons (*8*). In the cases for which only the *C. elegans* ortholog has been genetically studied, detailed analysis of the *C. elegans* mutant can suggest the function of the mammalian orthologs (Table 1). Thus, extrapolating from electron microscopic reconstruction of identified neurons and behavioral studies of the *C. elegans* mutant, the mammalian *unc-4* ortholog may regulate features of cholinergic motor neurons, such as connectivity (*9*).

Seven of the *C. elegans* homeobox genes are probable orthologs of the *Drosophila* and vertebrate *Hox* cluster, corresponding roughly to one *eve* (*vab-7*), two *AbdB* (on YAC Y75B8), two *Antp* (*egl-5* and *mab-5*), one *Scr* (*lin-39*), and one *Labial* (*ceh-13*) class homeobox genes (*7*). Mutations in the four of these genes that have been studied genetically affect patterning along the anterior-posterior axis as precedent from *Drosophila Hox* cluster genetics would predict (*7*). The two *AbdB* class genes were previously missed in *Hox* cluster molecular and genetic analyses; these genes are closely related and located within 5 kb of each other, suggesting a recent gene duplication and possible redundancy. The *C. elegans Hox* genes are located on the same chromosome but are distributed over 3 Mb, with thousands of intervening genes, unlike the tandemly arranged clusters in *Drosophila* or vertebrates. The *C. elegans Hox* cluster is also missing particular genes that are present in both the *Drosophila* and vertebrate *Hox* clusters. The simplicity and partial dispersal of the *C. elegans Hox* cluster, as well as the phylogenetic placement of the Nematoda to the same phylogenetic lineage as arthropods suggest that its *Hox* cluster may be a derived, deleted version rather than a primitive ancestral *Hox* cluster. There is no detectable *C. elegans Parahox* cluster, although the *caudal* ortholog, which constitutes one member of the Amphioxus *Parahox* cluster, is located on the same chromosome as the disintegrating worm *Hox* cluster.

The Polycomb (Pc) group of chromatin proteins have been implicated in repression of *Hox* cluster gene expression and heterochromatin formation in *Drosophila* and other animals (*10*). Many of the *Drosophila* Pc group that have mammalian orthologs are not present in the worm genome. Only *C. elegans* orthologs of the Pc group genes *enhancer of zeste* and *Esc* can be detected and mutations in these

genes, *mes-2* and *mes-6*, respectively, have defects in gametogenesis but not spatial patterning (*11*). Missing, for example, are orthologs of the *Drosophila* Pc group genes *Psc*, *Pc*, *Ph*, *Pcl*, and *Scm*, all of which detect probable mammalian orthologs. The phylogenetic placement of *C. elegans* in the Ecdysozoa suggests that most of these chromatin remodeling proteins have been lost in the *C. elegans* evolutionary lineage. Those Pc group genes that remain may no longer function in maintenance of *Hox* gene expression. In contrast, there are clear *C. elegans* orthologs of the *trithorax* class genes *ash1*, *ash2*, *Trx*, and *Brahma* that have been implicated in establishment of *Hox* gene expression in *Drosophila* (*10*).

Perhaps the partial dispersion of the *C. elegans Hox* cluster linkage and the loss of most Pc group genes are linked. For example, long range Pc class regulation of *Hox* gene chromatin structure may impose a genetic selection on the integrity of the *Hox* gene cluster that is so striking in the arthropod and chordate lineages. The relatively recent loss of this form of gene regulation in *C. elegans* may allow its *Hox* cluster to disperse. The presence of the *C. elegans Hox* genes on the same chromosome but not organized in tandem may represent a cluster in the process of disassembly after loss of the Pc genes. The model that ancestrally linked genes tend to initially diffuse apart over the same chromosome and then to other chromosomes is supported by many examples in the *C. elegans* genome of closely related genes localized to the same genetic region whereas more distantly related gene families tend to be localized to the same chromosome.

The orthologous sequence relationships of 51 of the 83 *C. elegans* homeobox genes and other vertebrate or invertebrate homeobox genes indicate probable conserved functions, but there are examples of orthologs that also have novel functions in particular species. The *Apterous*/*ttx-3*LIM homeobox gene orthologs function in the generation of thermoregulatory neural circuits in *C. elegans* and perhaps in *Drosophila* and vertebrates (*12*) and also regulate limb development in *Drosophila* and vertebrates (but not in *C. elegans*, which has no limb equivalent). It is probable that the neural developmental role is ancestral, and that in the arthropod and tetrapod lineages, these genes have acquired transcriptional regulatory sequences to broaden their function to the developing limb. Conversely, *C. elegans* carries probable orthologs of homeobox genes implicated in eye development across phylogeny (*barh1*, *eyeless*, *and sine oculis*), but *C. elegans* does not have eyes. Either these genes mediate events in the development of a modified *C. elegans* eye (conceivably an infrared sensing thermosensor), or these genes have been hijacked to another role. Thus the orthology of 51 *C. elegans* homeobox genes indicates only a subset of their possible functions.

There are 32 *C. elegans* homeobox genes that do not cluster in dendrograms with genes from other invertebrate or vertebrate databases; 16 of these genes can be classified into homeobox subtypes but 16 genes define novel subtypes. Missing orthologs of *C. elegans* genes in other species could be due to limited expression because in general homeobox genes have been sampled from cDNA rather than genomic libraries, or they could reflect an expansion of particular gene functions in *C. elegans*.

There are eight examples of *Drosophila*/vertebrate pairs of homeobox orthologs that are not detected in the worm genome (*13*), and more homeobox genes detected only in *Drosophila* or only in vertebrates that are not present in *C. elegans*. The missing *Drosophila* and vertebrate genes do not appear to mediate developmental events in a tissue not present in *C. elegans*. Assuming that the missing vertebrate/*Drosophila* orthologs are not in the estimated 1% of the *C. elegans* genome sequence that remains to be determined, they have probably been deleted from the *C. elegans* genome and may be missing more broadly in the phylogenetic lineage that includes *C. elegans*. Whatever the evolutionary history of their loss or divergence, *C. elegans* can develop without these genes.

Other transcription factor genes implicated in developmental control. Transcription factors of the basic helix loop helix, bZip, zinc finger, forkhead, ETS, and T-box families (each of which can be recognized by a distinct protein sequence motif) have been implicated by genetic and molecular analysis in cell type specification in a variety of systems (Table 1). There are multiple *C. elegans* genes in each family, many of which are orthologous to mammalian genes. About 10 to 25% of the genes in these families have emerged from *C. elegans* genetic analysis of cell fate specification and signaling suggesting that much analogous developmental control remains to be studied. A number of the *C. elegans* genes in each family have no clear vertebrate or other invertebrate orthologs, and particular conserved subclasses in each family detect no *C. elegans* orthologs, suggesting that the families have expanded and diverged (or have contracted differentially) in each phylogenetic lineage. For example, a striking feature of the 14-member *C. elegans* T box family is that it is missing a *C. elegans* ortholog of Brachyury, which regulates the development of the chordate notochord and *Drosophila* hindgut differentiation (*14*). Thus, the *C. elegans* genome sequence does not reveal a possible ancestral function of a key chordate regulator; rather, it suggests that metazoan development is possible without this gene.

The most dramatic expansion is seen in the *C. elegans* nuclear hormone receptor (NHR) gene family. The NHRs are DNA-binding proteins that bear a variety of ligand-binding domains for small molecules and regulate diverse developmental and physiological processes. There has been a huge proliferation of NHR genes in the *C. elegans* genome (*15*). Only 12 of the 235 genes detect probable mammalian or *Drosophila* orthologs (*15*). Many of the orthologous *C. elegans* NHR genes and two of the six genetically analyzed *C. elegans* NHR genes that affect molting, *daf-12* and *nhr-23*, detect components

Table 1. Transcription factor genes.

Gene family	Genes identified		Orthologs	
	By sequence*	Genetically†	In *C. elegans*‡	Not in *C. elegans*§
Homeobox genes¶	83	24	51	11
HOX cluster	7	5	7	4
PAX	2	1	2	1
POU	3	3	1	0
LIM	7	6	6	0
Prd-type	14	4	9	0
Nkx2	4	2	3	0
SIX	5	0	3	0
TALE	4	1	3	0
CUT	5	0	1	0
ZFH:	2	0	2	0
BarH1:	2	0	2	0
Orthol. pairs	12	3	12	–
New class	16	0	0	6
Paired box	5	2	5	0
T-box	17	2	1	3
bHLH	24	6	11	1
bZip	19	2	10	1
Forkhead	15	4	6	5
ETS	10	1	6	2
NHR	235	6	12	ND
Polycomb-group	2	2	2	7
Trithorax	4	0	4	2

*A list of the gene names can be found at www.sciencemag.org/feature/data/985556.shl. One gene, the zfh-2 ortholog zc123.2, has 3 homeodomains. †This category includes genes that have been positionally cloned on the basis of a mutant phenotype and genes that have been knocked-out by gene-targeting methods. ‡Phylogenetic trees on which the orthologous assignment is based can be found at www.sciencemag.org/feature/data/985556.shl. §This category includes only those genes that have vertebrate/invertebrate orthologs, but no *C. elegans* ortholog. ¶Classes with at least two *C. elegans* genes are listed separately. Single *C. elegans* genes that have invertebrate/vertebrate orthologs fall into "Orthol. pairs," whereas genes with no orthologs constitute a "New class." ND, not determined.

of the *Drosophila* ecdysone response cascade, a provocative result for a molting animal (*16*). However, the *sex-1* NHR gene is also orthologous to *Drosophila* ecdysone response pathway genes, but has been genetically implicated in X chromosome counting in sex determination and dosage compensation rather than molting (*17*). Notably missing from *C. elegans* is an ortholog of the insect ecdysone receptor, suggesting that a conserved hormonal response pathway that is triggered by ecdysone in insects may be triggered by some other hormone in the *Caenorhabditae*. The remaining 223 *C. elegans* NHR genes are not represented in other partial genome sequences, suggesting that many are nematode-specific. Consistent with a rapid and recent expansion of this gene family in *C. elegans*, many of NHR genes map to one chromosome, suggesting that these genes multiplied in recent evolutionary history and have not yet drifted to other genetic regions (*15*).

Conservation in binding sites for transcription factors. The orthologous relationships among so many transcription factors, and their assignment to functionally related genetic cascades suggests that these transcription factors may bind to and regulate common targets across phylogeny. In fact, this has been shown to be the case in muscle development from *C. elegans* to chordates, where the bHLH transcription factor Twist regulates the expression of NK-class homeodomain transcription factors in the muscle regulatory cascade (*18*). However, whereas the regulatory regions of the NK genes bearing the bHLH protein binding sites are clearly conserved between *D. virilis* and *D. melanogaster* and between *C. elegans* and *C. briggsae* (species pairs phylogenetically separated by 10^7 years) they are only weakly conserved between *C. elegans* and *Drosophila*, separated by a 100-fold longer period. The genome sequences of species divergent between 10^7 and 10^8 years from *C. elegans*, *Drosophila*, and vertebrates may need to be determined to allow simple informatic identification of regulatory binding sites.

There are examples of enhancer function conserved across species, most notably between autoregulatory elements of the *Hox* gene cluster of arthropods and vertebrates (*19*). Because of the precision with which expression patterns can be determined in *C. elegans* and correlated with the expression patterns of transcription factors, future developmental genomics may test expression patterns in *C. elegans* of candidate enhancer elements from other species identified from informatic analysis of genome sequences.

Signaling Pathways

Consistent with the findings from classical embryology that key signaling centers control development, a variety of growth factor signaling molecules have been genetically shown to regulate *C. elegans* development. More members of these families are detected in the genome sequence and are likely to regulate other developmental events.

TGF-β pathways. Members of the TGF-β family of extracellular signals, as well as the receptor and signaling pathways downstream, have been implicated in early patterning and physiological regulation of both vertebrates and invertebrates (*20*). The *C. elegans* TGF-β signaling cascade is the paramount example of a signal transduction superfamily nearly saturated by genetic analysis (*20*). There are four TGF-β ligand family members in the *C. elegans* genome, and three of these have been genetically analyzed. The genome sequence reveals two type I receptors, one type II receptor, and six *Smad* proteins that transduce signals from the receptors to the nucleus, and the functions of all of these genes have been genetically studied and ordered into two pathways.

The four TGF-β–like ligands are quite divergent from those in other species but one of them, DBL-1, clusters with *Drosophila dpp*/vertebrate *BMP-4* (*21*). DBL-1 mediates body size determination from its site of expression in the ventral cord (*21*). Thus even though it is a probable ortholog by informatic classification, DBL-1 functions more as an endocrine signal of the activin/inhibin type than an early patterning gene like *dpp* or *BMP-4*. DAF-7 mediates metabolic and diapause control in a neuroendocrine fashion from its expression in one or a few neurons (*22*). UNC-129 mediates neural pathfinding (*23*). The fourth TGF-β gene has not been studied genetically but is most closely related to *BMP-7*/*Drosophila 60A*.

The genome sequence reveals only two pathways by which these four TGF-β ligands may signal. There are two type I receptor kinases for these ligands: The DAF-1 type I receptor kinase has been genetically shown to transduce DAF-7 TGF-β signals (*24*) and the SMA-6 type I receptor kinase acts in DBL-1 body size signaling (*25*). The DAF-4 type II receptor is the only type II receptor kinase and, consistent with this genomic analysis, transduces both DBL-1 and DAF-7 signaling (*24*). These type I and type II receptors detect homologous proteins in vertebrate and other invertebrate databases, but do not cluster obviously with particular family members to allow simple orthology assignments.

Of the six *Smad*s revealed by the *C. elegans* genome sequence, three have been genetically implicated in DAF-7 signaling, and three others have been genetically implicated in DBL-1 signaling (*25*, *26*). The genomic ratios of three *Smad*s per cognate receptor pair and the biochemical finding that mammalian *Smad* proteins form trimers upon receptor activation (*27*) suggest that the three *Smad* proteins in each pathway form heterotrimers to propagate TGF-β signals to downstream genes. Because these are the only TGF-β receptor and *Smad* genes in the genome, the UNC-129 and other TGF-β ligands are also likely to couple via these transduction pathways. The genome analysis leaves no room for other canonical TGF-β receptors or Smads.

Consistent with the *C. elegans* TGF-β signaling ligands acting in a neuroendocrine manner, no homologs are present in the worm genome of vertebrate *chordin*/*Drosophila sog*, which bind to TGF-β ligands to confer short range gradients for patterning (*28*). However, there is a probable *C. elegans* ortholog of the *tolloid*/*BMP-1* metalloprotease gene that processes *sog* and chordin. Perhaps *the tolloid* class of proteases have other functions besides sog/chordin processing.

Receptor tyrosine kinase pathways. Receptor tyrosine kinases (RTKs) were originally revealed as regulatory genes from their action in growth factor signaling and oncogenic pathways, but have also been shown to mediate patterning events in both vertebrates and invertebrates (*20*). There are 28 *C. elegans* RTK genes, 11 of which correspond to probable orthologs of other vertebrate and other invertebrate RTK genes. Mutations have been identified in four of the RTK genes, all of which act in genetic pathways that support their assignment to orthologous pairs: DAF-2 is an insulin/IGF-I receptor ortholog that also mediates metabolic and growth control (*29*); EGL-15 is an FGF receptor ortholog that mediates mesodermal migration signaling from the FGF-related ligand EGL-17 (*30*); VAB-1 is the probable EPH receptor kinase ortholog that mediates head and tail neural and hypodermal patterning (*31*); LET-23 is the EGF receptor ortholog for the epidermal patterning TGF-α–related ligand LIN-3 (*32*).

Given how extensively RTK genes have been sought in mammalian and other invertebrates, the two families that are specific to *C. elegans* may actually be unique to this phylogenetic lineage. In fact, one of the *C. elegans*-specific RTK families appears to have undergone a recent expansion to 11 members; the members are all located within 500 kb of each other, as if they have diffused by inversion from the point of multiplication. Two of these family members, *kin-15* and *kin-16*, are expressed in hypodermal cells (*33*), and the 11-gene *kin-15/16* family is interspersed with chitinase genes, suggesting perhaps an involvement in epidermal fungal resistance.

There are also missing *C. elegans* RTK genes that are present in *Drosophila* and vertebrates: For example, there are no nerve growth factor/*trk* receptor or PDGF/FLK receptor genes. It is possible that the small size or short lifespan of *C. elegans* supercedes the former requirement for neuronal survival factors such as NGF.

Table 2. Growth factor signaling genes.

Gene family	Genes identified	
	By sequence*	Genetically
TGF-β signaling pathway		
TGF-β–like ligands	4	2
TGF-β–like receptors	3	3
Smad proteins	6	6
Wg/Wnt signaling pathway		
Wnt-like ligands	5	3
Fz-like receptors	4	2
Signal transduction components:		
APC	1	1
dsh	3	1
GSK-3	2	1
β-catenin	3	3
LEF/TCF	1	1
Other		
Porc	1	1
Axin	0	0
GBP	0	0
lin-12/*Notch* signaling pathway		
lag-2–type ligand	4	2
lin-12–type receptor	2	2
su(H)	1	1
Groucho	1	1
Neuralized	1	0
Sno	1	0
Numb	1	0
Prospero	1	0
Kuzbanian	1	1
Presenilin	3	3
Cdc-4	1	1
Deltex	0	0
Mastermind	0	0
Hairless	0	0
Scabrous	0	0
Big brain	0	0
Fringe	0	0
Receptor tyrosine kinase signaling pathways		
Receptors†	28	6
Signal transduction components:		
Grb2	1	1
Nck	1	0
Crk	1	0
Cbl	1	1
Cnk	1	0
IRS-1	0	0
Shp2/csw	1	1
Sos	1	0
Ras	1	1
GAP	2	2
Raf	1	1
Ksr	1	1
Sur-8	1	1
Akt	2	2
PI3K:		
p110	1	1
p85-like	1	1
Mek‡	2	2
Mapk	1	1
PTEN	1	1
Cytoplasmic tyrosine kinases		
Src	1	1
Abl	1	0
Fak	1	1
Fes/Fer	19	0
Syk/Zap70	0	0
Toll/IL1 receptor signaling pathway		
Toll-like receptor§	0	0
Pelle kinase	1	0
NFκβ/*dorsal*, Iκβ	0	0
Cytokine signaling pathways		
Cytokine receptors (gp130, βc, γc)	0	0
Jak kinase	0	0
Stat	0	0
Hedgehog signaling pathway		
Hedgehog-like ligand	0	0
smo receptor	0	0
Patched receptor	1	1
Fused kinase	0	0
Protein kinase A	1	0
ci/gli tsk factor	1	1
Tout-velu	1	0

*A list of the gene names can be found at www.sciencemag.org/feature/data/985556.shl. †See figure at www.sciencemag.org/feature/data/985556.shl. ‡There are many more MEK and MAPK-like genes. Only one each, however—being the best homologs of vertebrate MEK and MAPK, respectively—has been implicated in RTK signaling. §There are many LRR-containing receptors, none of which, however, is closely related to Toll/IL1R.

The orthologous relationship among a number of *C. elegans* and vertebrate RTK genes allows the attributes of the known ligands of the well-characterized receptors to be used to search the *C. elegans* genome database for ligands. As a general rule, the ligands are much more difficult to detect because they are small and diverge quickly in evolution, so our ligand family sizes may be underestimates. Because there is only one member in the genome of the RTK subtypes discussed below, it is probable that any ligand family members couple via their corresponding RTK only. For example, 10 insulin-like ligands for the DAF-2 RTK have been reported (*34*). Much of the expansion of this *C. elegans* insulin family appears to be recent because many of the genes are clustered (*34*). Even though these genes are recently duplicated, their protein sequences are highly diverged, suggesting extraordinary selection. These ligands may be expressed in distinct sets of cells or regulated differentially by distinct environmental inputs. The recent generation of this complexity evinced by the genetic clusters suggests that such regulatory complexity may have been acquired in a saltatory fashion.

In the other RTK families, there are two FGF-family ligands for the EGL-15 receptor. There are four ephrin-related *C. elegans* genes that may correspond to ligands for VAB-1. However, no other ligands for the LET-23 EGF type receptor emerge from comparing the *C. elegans* genome with known ligands for this receptor family from worms, flies, and mammals, suggesting that these particular growth factors diverge quickly in evolution.

Downstream of these receptors, signal transduction proteins such as adaptor proteins that couple to GTP regulated *ras* signaling and kinases have been identified by genetic analysis. Surprisingly, each signaling family does not reveal a constellation of related proteins: There is just one gene family member in the *C. elegans* genome for most of the signal transduction proteins listed in Table 2. Some of the cytoplasmic ser/thr kinases detected in the genome are likely to act downstream of these receptors, as has been shown for particular kinases downstream of LET-23 and DAF-2. It is striking, however, how few cytoplasmic kinases have been identified by developmental genetics in *C. elegans*. Whereas one-quarter of the RTKs were

identified by classical genetics, only a few of the more than 300 cytoplasmic kinases have emerged from genetic analysis. The probable explanation is that multiple ser/thr kinase pathways emerge from the receptors so that mutation in any one kinase causes subtle effects not easily recognized by the geneticist.

Wnt Pathways

Growth factors and signaling pathways of the *Wnt/wingless* class have been implicated in patterning in both vertebrates and invertebrates (*35*). There are four probable *C. elegans Wnt* pathways revealed in the genome sequence that may utilize some common downstream signaling components and some specific components (Table 2). Three of the *Wnt* pathways have been studied genetically and shown to each control distinct oriented cell divisions or cell migrations; thus the *Wnt* signaling cascade is almost fully revealed by genetic analysis (*36*). Of 12 *C. elegans* members of the LEF/TCF family of *Wnt* transcriptional outputs, one gene has been implicated by genetic analysis as an output of a *Wnt* signaling pathway (*36*). Of three β-catenin genes that encode bifunctional proteins that transduce *Wnt* signals to the nucleus as well as play structural roles in adherens junctions, all have emerged from genetic analysis (*36–38*). One of the β-catenin genes acts in a *Wnt* pathway, but the other two have not been assigned to *Wnt* pathways, and one (*hmp-2*) may play a more structural role in morphogenetic movements (*36*, *37*).

lin-12/Notch Pathways

The *lin-12/Notch* receptor signaling pathways mediate cell-cell signaling to pattern equipotential cells in a variety of systems (*39*). The genome sequence reveals only two *C. elegans lin-12/Notch*–type signaling pathways and both of these have been extensively explored by genetic analysis. There are four *Delta/jagged* ligand genes, two of which have been shown genetically to act in the pathways. There are two *lin-12*/Notch receptor homologs that function redundantly for some signals and independently for others. The two receptor genes are also closely linked, suggesting a recent gene duplication and partial functional divergence. The specificity of their function is mainly due to distinct patterns of expression, suggesting that it is the evolution of transcriptional regulatory diversity that has generated the diverse functions of these recently duplicated genes (*39*). All four ligands are likely to couple via these receptors, perhaps in distinct tissues. A *C. elegans Suppressor of Hairless* ortholog is likely to be the major transcriptional output of both receptors. Also detected in the genome sequence but not yet implicated in *lin-12/Notch* signaling by genetic analysis are probable orthologs of a variety of *Notch*-coupled proteins from *Drosophila* (Table 2). Other genes implicated in *Drosophila Notch* signaling are missing from the *C. elegans* genome (Table 2).

Cell Death Pathway

Programmed cell death is a major feature of vertebrate and invertebrate development. Molecular genetic analysis of *C. elegans* genes that either induce or prevent programmed cell death revealed a regulatory pathway that detects mammalian homologs that are now understood to regulate cell death as well (*40*). Thus, the *C. elegans* genome sequence is an opportunity to find other possible cell death regulators. The CED-9 cell death–protecting protein is the only Bcl-2 homolog in the *C. elegans* genome. The EGL-1 cell death–promoting protein bears one of the four *bcl* family motifs; there is a second *C. elegans* BH domain gene that has not been implicated in cell death. The ICE protease/caspase gene *ced-3* is one of three *C. elegans* caspase family members; the other genes have not been implicated in cell death. The gene *ced-4* is related to mammalian *apaf-1* and is the only relative in the *C. elegans* genome. Two out of three molecularly analyzed *C. elegans* cell death–related engulfment gene products have mammalian orthologs: the DOCK180 protein CED-5 and the ABC transporter CED-7, whereas CED-6 is a novel protein. Only in the case of *ced-7* are there other members of the gene family in *C. elegans*, but these are likely to be general small molecule transporters. There are no *C. elegans* genes related to the *Drosophila* cell death regulatory genes *reaper*, *grim*, and *hid*, or the mammalian cell death genes *fas* or its receptor, TNF-α or its receptor, or FADD and RIP. There are two *C. elegans* genes related to the cell regulatory gene IAP but these have not been genetically studied.

Developmental Pathways That Are Divergent in *C. elegans*

Although the *Notch/lin-12* and *wingless* signaling pathways mediate the interactions between *C. elegans* blastomeres after the first few cell divisions, it appears that novel molecules identified by genetic analysis have evolved to specify the generation of the first blastomeres (*41*). Many of the genes have motifs that reveal their biochemical function, but most detect no orthologs. For example, *mex-1* and *pie-1* encode divergent Zn finger proteins that are asymmetrically expressed in early blastomeres and may differentially bind maternal mRNAs to specify blastomere cell fates (*41*). Similarly, the asymmetrically expressed *skn-1* encodes a divergent member of the bZIP transcription factor class that specifies blastomere fates (*41*).

Upstream of these transcription and translation factors, the *par* genes regulate the asymmetry of early blastomere cell divisions (*41*). Many of the proteins encoded by the *par* genes are asymmetrically localized in the early embryo to mediate the asymmetric activation or sorting of the translation and transcriptional regulatory proteins. Most of the *par* genes do not detect orthologs in other species.

The genes that regulate *C. elegans* early development may be specialized for early blastomere patterning in this phylogenetic lineage. *C. elegans* embryonic development is cellular, in contrast to *Drosophila* syncytial development. The cellular form of early development is more like that in vertebrates, so it is possible that orthology to vertebrate genes will emerge. On the other hand, the asymmetric blastomere cleavages of *C. elegans* are more similar to the spiral cleavages of many invertebrate species; orthologs of *C. elegans* genes involved in this process may emerge from analysis of these other species.

Another major focus of *C. elegans* developmental genetics has been a dissection of how the temporal dimension of development is regulated (*42*). Most of the *C. elegans* genes that regulate temporal patterning detect no clear *Drosophila* or chordate orthologs. Thus it is possible that the temporal patterning genes of *C. elegans* identify components of a pathway that may be limited, for example, to particular molting species or an even narrower phylogenetic distribution. The product of the heterochronic gene *lin-4* is a 25-nt regulatory RNA that controls the translation of other heterochronic genes (*42*). Although this gene is clearly conserved in the *Caenorhabditae*, it cannot be detected in other genomes. Regulatory RNA genes such as *lin-4* escape detection by gene-finding programs, but will most likely emerge from sister genome comparisons akin to those between *C. elegans* and *C. briggsae* (as well as by genetics, as in the case of *lin-4*).

Missing Pathways in *C. elegans*

Until complete genome sequences became available, one could never conclude that a pathway was missing in any organism; only that researchers had failed to find it. The nearly complete *C. elegans* genome sequence allows us to enumerate, for the first time in any animal, missing developmental control genes that are broadly represented in animal phylogeny.

There are two classes of missing *C. elegans* gene superfamily members. Missing *C. elegans* orthologs that are present in both chordates and other invertebrates have probably been deleted in the phylogenetic lineage that leads to *C. elegans*; genes found in such divergent animal clades would be expected to be present in another Ecdysozoan, *C. elegans*. The missing *C. elegans* gene classes that

have been defined by genes in a single phylogenetic branch may correspond to inventions in one phylogenetic lineage rather than deletions in the *C. elegans* phylogenetic lineage.

The most striking missing (or highly modified) pathway is hedgehog (*hh*) signaling. Even though a variety of invertebrate and vertebrate species specify significant body pattern elements by *hh* signaling (*hh* ligand, the *smoothened* and *patched* receptors, the fused and PKA kinases, the *ci/gli* transcriptional output), *C. elegans* is missing the *hh* ligand, the smoothened receptor, and the *fused* kinase of the pathway. These proteins are conserved over large regions between *Drosophila* and many vertebrate species, so it is not likely that the *C. elegans* orthologs would be unrecognizable (*43*).

The assertion that any one gene is missing in *C. elegans* is subject to the caveat that gaps in the genome sequence still exist that are estimated to contain up to 1% of the total genes. So for each gene declared missing in action, it is really a probability of 99% that it is missing. In the case of *hh*, the missing three pathway genes (*hh*, *smoothened*, and *fused*) leaves an even more remote chance that all three genes will show up in the eventual complete *C. elegans* genome sequence.

Even if there is no *C. elegans hh* ortholog, the probable ortholog of *patched* and a second *patched*-like receptor gene, both of which bear conserved intracellular and extracellular regions, are present in the *C. elegans* genome. Neither of these genes have yet been implicated in any genetic pathway but P. Kuwabara has found that both *patched* family members are essential genes (*44*), suggesting that the *patched* signaling pathway without a *hh* ligand or *smoothened* receptor continues to function in *C. elegans*.

The *C. elegans* ortholog of the Zn finger transcriptional output of *hh* signaling, *ci/gli*, emerged from molecular genetic analysis of the sex determination pathway as the gene *tra-1* (*45*). The sex determination pathway has been genetically studied in great detail (*45*), and no other member of the remaining *hh* signaling pathway has been implicated in control of *tra-1*. However a signaling cascade is coupled to *tra-1*, and the receptor in that cascade, TRA-2, bears very weak sequence similarity to the *patched* family of receptors (*46*). The HER-1 ligand for the TRA-2 receptor, however, has no sequence similarity to *hh*. The TRA-3 protease has been implicated in the pathway that includes the receptor and TRA-1, and the *ci* ortholog is proteolytically cleaved upon activation of *hh* signaling (*47*). Sex determination pathways may be quite plastic in evolution. Thus it is reasonable to see a possible orphan from a lost *hh* signaling pathway, *tra-1*, reassigned to sex determination. The extant *patched* "ortholog" may have been similarly reassigned, or the *patched* gene family across phylogeny may in fact function in more than just *hh* signaling, and only the other function remains in *C. elegans*. At the very least, we must conclude that metazoan development is possible without *hh* patterning.

Another example of partial pathway deletion is the Toll/IL1 immunity signaling pathway. Although there is no clear *C. elegans* ortholog of the Toll/IL1 receptor, there are a few members of this receptor family. There is a probable *pelle* kinase ortholog, but no *rel* or *dorsal* homolog, which are the transcriptional outputs of these pathways in vertebrates and *Drosophila*, respectively (*48*). The ancient function of the Toll/IL1 receptors in immunity may be conserved but its transcriptional outputs are not.

The JAK/STAT signaling pathway is even more vestigial in *C. elegans*. Whereas *Drosophila* and vertebrates have a cytokine receptor to JAK kinase to STAT transcription factor signaling pathway that regulates hematopoiesis as well as other growth and physiological signals (*49*), no receptor homologs, no JAK kinase homologs, and only a very distant partial STAT gene are present in *C. elegans*.

Although *C. elegans* may be the first animal for which we can list what is missing, the deletion of almost universal orthologs may be a general phenomenon. For example, the genome of the ancestor to nematodes, arthropods, and chordates may have been more complex, but may have been trimmed in distinct steps in each of the phylogenetic lineages. In such a case, when complete genome sequences emerge from arthropods and vertebrates, distinct sets of missing genes will be found there as well. It is possible that the selection of small genomes for sequencing projects enriches for species that have trimmed genomes. On the other hand, *C. elegans* may have a particularly trimmed down genome so that a variety of genes have been lost somewhere in the phylogenetic lineage to *C. elegans* but not in other lineages. Searching for the missing set of phyletically general genes in the Ecdysozoan and Nematoda lineages may illuminate how genes are lost or reassigned in *C. elegans*. If the genes are found in species more closely related to *C. elegans* than *Drosophila* or chordates, the deletion of this pathway in the *C. elegans* lineage will be supported.

The Utility of the Genome Sequence in Present-Day *C. elegans* Developmental Genetic Analysis

The previous molecular analyses of approximately 300 genetically identified *C. elegans* genes has generated a correlated physical and genetic map that speeds the positional cloning of subsequent genetic loci. The 1300 genetic loci (about 6% of the total genes) that have not yet been molecularly analyzed will fuel future genetic analysis of developmental control. Genes in a broad genetic region can be elevated to candidacy for a mutant being mapped, on the basis of current molecular knowledge of the pathway or the mutant phenotype. Candidate genes can be tested without laborious fine genetic mapping, for example, by complementation with cloned DNA segments or by sequencing of mutant alleles. The recent advent of double-stranded, RNA-mediated gene inactivation has revolutionized our ability to inactivate candidate genes in a region and assess the phenotypic consequences (*50*).

The genome sequence also speeds other genetic manipulations. Primers for amplification by the polymerase chain reaction can be designed from the genome sequence to allow simple amplification and genetic transformation of particular genetic regions, so that gene dosage can be increased in wild type or pathway mutants to search for high gene dosage suppression or enhancement of phenotypes. Gene fusions to the green fluorescent protein have been used extensively to reveal gene expression patterns and as molecular phenotypes with which to characterize or screen for other mutants (*1*). Hybrid genes can now be constructed to misexpress genes in novel cell types to prove necessity or sufficiency of gene activities, or express genes from other organisms to prove orthology. Gene activities can be disrupted by efficient sib-screening technologies (*51*).

Saturation genetic analysis of those pathways that already show clear orthology to mammalian pathways are promising avenues for practical impact on human health; for example, by detecting orthologs in humans of each new gene revealed by genetics. There are also cases in which the detection of orthology does not just add a new gene; it fuses two pathways that were previously unrelated. For example, saturation genetic analysis of *lin-12* receptor signaling revealed that a *C. elegans* ortholog of the human PS1 multipass transmembrane protein, which had been implicated by human pedigree analysis in predisposition to Alzheimer's disease, regulates the *lin-12* function (*39*). Thus the mammalian PS1 gene that had been placed in neural degeneration pathway was now also placed in a *Notch*-like signaling pathway, and vice versa, fusing what were considered disparate fields. This research suddenly thrust the well-developed *lin-12* signaling pathway into candidacy for regulation of neural degeneration in Alzheimer's disease.

Conclusion

We have been in a gene discovery era that celebrates the listing of universals. It appears that about half of the *C. elegans* genes belong to gene families that are broadly distributed across animal phylogeny

and about half will have a more restricted phylogenetic distribution. The set of universal developmental control genes, such as the *ras*, presenilin, insulin, and cell death pathways, promise to reveal mechanisms that are important to human development and health.

Although animals have much in common, worms are not puppies and humans are not sea cucumbers. An outstanding question is whether the obvious differences between animal forms and functions are caused by changes in regulation of the universal genes, or caused by differences in the complement of related or novel regulatory genes that might be more specific to particular phylogenetic lineages, or both.

The dendrograms of each gene family are the historical record of multiple gene duplications and divergences of function over the course of evolution. Some of this gene duplication and divergence occurred in pre-Cambrian times and have been inherited by much of animal phylogeny. These are the orthologous genes that serve conserved roles in nearly all animals. The members of the gene families that do not detect orthologs may be recent inventions in particular phylogenetic lineages, or may have been lost in particular phylogenetic lineages. This is the evidence of the ongoing invention in these gene families.

Perhaps the most striking feature of these trees is how genes on most of the branches, whether unique to particular phyla or universal, have already been shown by genetic analysis (mostly in flies and worms) to function in particular developmental events. Thus, the ancient duplication and divergence of these gene families that generated the deeply orthologous members as well as the more recent phylogenetically restricted diversification have generated essential functions.

Even within the universal set of broadly orthologous genes, novelty emerges. There are many examples of orthologs taking on new roles in evolution, as if a core function has been conserved but additional functions have evolved in particular phylogenetic lineages. It is essential to establish which gene function of an ortholog is more phyletically general and thus primitive and which is derived in the comparisons of orthologs across phylogeny. Just as the universals and peculiarities of *Hox* gene function have been a major object of recent phylogenetic comparisons, we can expect many such orthologous pathways to be so analyzed.

From the current state of analysis of these genes, it is already clear that new gene functions are generated both by acquisition of novel transcriptional regulatory domains—in most cases, the diverse members of gene families are expressed in distinct cell types or times or places—as well as by the divergence of the protein sequences to interact with distinct other molecules. Comparative genomics may reveal where the regulatory regions come from in this reassignment of function. It would not be surprising to find that transposons mediate the construction of recombinant regulatory elements, perhaps even by horizontal gene flow between species.

The plasticity of gene orthologs in phylogeny may provide the best opportunity to view the acquisition of new functions. However, it will also be interesting to view the trajectory of gene family members that are specific to particular phylogenetic lineages: Where did they come from? Have they been deleted in species that are missing them, or have they evolved quickly under selection in some phylogenetic lineages by point mutation or recombination between family members?

Comparative analysis of other genomes will drive this analysis. A comparison of the complete genome sequences of key animals from each phylogenetic lineage, defined by an approximate 10^8-year phylogenetic distance from the cardinal genome points (*C. elegans*, *Drosophila*, and human) would reveal many of the commonalities and distinctions. The distribution of particular subtypes in gene families and determination of their expression patterns may give initial hints of their evolution and their functions. An analysis of the expression patterns of a selection of universal and phylogenetically restricted developmental control genes in a broad phylogenetic sampling of developing animals, would reveal which aspects of their regulation are universal and which are unique to particular species. In this way, we may view the evolutionary generation of diversity from the basic building blocks we now understand.

References and Notes

1. W. B. Wood, Ed., *The Nematode* C. elegans (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1988); D. L. Riddle *et al.*, Eds., C. elegans *II* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1997).
2. We classify gene superfamilies mainly based on phylogenetic analyses that are presented on the Web page www.sciencemag.org/feature/data/985556.shl. Routinely, searches were first performed using BLASTP on the Wormpep dataset posted on the *C. elegans* genome blast server www.sanger.ac.uk/projects/c_elegans/blast_server.shtml (which at the time of this analysis was Wormpep15 plus a number of unannotated sequences now included in Wormpep16) and afterward confirmed using TBLASTN on the genomic sequence dataset on the *C. elegans* blast server. The probable ortholog of a *C. elegans* gene is the most closely related vertebrate or other invertebrate protein that by BLAST and/or dendrogram analysis (using the neighbor-joining method on a distance matrix created with the Kimura protein distance algorithm) is significantly more closely related than the next most closely related member of the gene superfamily. General information about the datasets and annotations can be found at: http://www.sciencemag.org/feature/data/c-elegans.shl.
3. K. Kornfeld, *Trends Genet.* **13**, 55 (1997); M. Hengartner and H. Horvitz, *Philos. Trans. R. Soc. London Ser. B* **345**, 243 (1994); S. Ogg *et al.*, *Nature* **389**, 994 (1997).
4. A. Aguinaldo, *Nature* **387**, 489(1997); A. Mushegian *et al.*, *Genome Res.* **8**, 590 (1998). Confirmation of this phylogeny comes from sequence comparisons by Mushegian *et al.* of orthologous protein sequences from vertebrates and other invertebrates to *C. elegans* proteins that have a rate of high amino acid substitution compared with those having low substitution rates: The slowly evolving set of *C. elegans* proteins cluster with their orthologs from molting animals, whereas the proteins not under so much selection appear to branch earlier from all other animal lineages.
5. W. Gehring, M. Affolter, T. Burglin, *Annu. Rev. Biochem.* **63**, 487 (1994).
6. T. Burglin *et al.*, *Nature* **341**, 239 (1989).
7. B. Wang *et al.*, *Cell* **74**, 29 (1993); C. Wittmann *et al.*, *Development* **124**, 4193 (1997); Y. Jin *et al.*, *Nature* **372**, 780 (1994).
8. M. Chalfie and J. Sulston, *Dev. Biol.* **82**, 358 (1981); L. Erkman *et al.*, *Nature* **381**, 603 (1996).
9. D. Miller *et al.*, *Nature* **355**, 841 (1992); J. G. White *et al. ibid.*, p. 838.
10. J. Simon, *Curr. Opin. Cell Biol.* **7**, 376 (1995).
11. R. Holdeman *et al.*, *Development* **124**, 2457 (1998); I. Korf *et al.*, *ibid.* **125**, 2469 (1998).
12. O. Hobert *et al.*, *Neuron* **19**, 345 (1997); C. Rodriguez-Esteban *et al.*, *Development* **125**, 3925 (1998); T. Klein *et al.*, *Curr. Biol.* **8**, 417 (1998).
13. The *C. elegans* homeobox genes orthologous to hlx, gbx/unplugged, mox/buttonless, otp, lbx/ladybird, pdx1, gsx, and tlx are not present in the genome sequence.
14. V. E Papaioannou and L. M Silver, *Bioessays* **20**, 9 (1998).
15. A. Sluder and C. Maina, *Genome Res.*, in press.
16. M. Kostrouchova *et al.*, *Development* **125**, 1617 (1998); A. Antebi *et al.*, *ibid.*, p. 1191.
17. I. Carmi *et al.*, *Nature* **396**, 168 (1998).
18. B. Harfe *et al.*, *Genes Dev.* **12**, 2623 (1998); Z. Yin *et al.*, *Development* **129**, 4971 (1997).
19. A. Awgulewitsch and D. Jacobs, *Nature* **358**, 341 (1992).
20. J. Smith, *Curr. Opin. Cell. Biol.* **7**, 856 (1995); D. L. Riddle and P. Alberts, in C. elegans *II*, D. L. Riddle *et al.*, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1997), p. 739.
21. W. Wood, personal communication.
22. P. F. Ren *et al.*, *Science* **274**, 1389 (1997).
23. A. Colavita *et al.*, *ibid.* **281**, 706 (1998).
24. M. Estevez *et al.*, *Nature* **365**, 644 (1993).
25. R. Padgett, personal communication.
26. G. I. Patterson *et al.*, *Genes Dev.* **11**, 2679 (1997); C. Savage *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 790 (1996); T. Inoue and J. Thomas, personal communication; D. L. Riddle, personal communication. One other divergent *Smad*, bearing only one of the conserved domains of the gene family, is also present but has not been genetically studied.
27. Y. Shi *et al.*, *Nature* **388**, 87 (1997).
28. G. Marques *et al.*, *Cell* **91**, 417 (1997).
29. K. Kimura *et al.*, *Science* **277**, 942 (1997).
30. R. D Burdine, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 2433 (1997).
31. S. E. George, *Cell* **92**, 633 (1998).
32. P. S. Kayne and P. W. Sternberg, *Curr. Opin. Genet. Dev.* **5**, 38 (1995).
33. W. P. Morgan and I. Greenwald, *Mol. Cell. Biol.* **13**, 7133 (1993).
34. L. Duret *et al.*, *Genome Res.* **8**, 348 (1998).
35. K. Cadigan and R. Nusse, *Genes Dev.* **11**, 3286 (1997).
36. C. Thorpe *et al.*, *Cell* **90**, 695 (1997); R. Lin *et al.*, *ibid.* **83**, 599 (1995); C. Rocheleau *et al.*, *ibid.* **90**, 707 (1997).
37. M. Costa *et al.*, *J. Cell Biol.* **141**, 297 (1998).
38. D. Eisenmann *et al.*, *Development* **125**, 3667 (1998).

39. I. Greenwald, *Genes Dev.* **12**, 1751 (1998); D. Levitan and I. Greenwald, *Nature* **377**, 351 (1995); X. Li and I. Greenwald, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 12204 (1997).
40. M. Hengartner, in C. elegans *II*, D. L. Riddle *et al.*, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1997), pp. 383–416.
41. K. Kemphues and S. Strome, in *ibid.*, pp. 335–360; R. Schnabel and J. R. Priess, in *ibid.*, pp. 361–382; S. Guo and K. Kemphues, *Curr. Opin. Genet. Dev.* **6**, 408 (1996).
42. F. Slack and G. Ruvkun, *Annu. Rev. Genet.* **31**, 611 (1998,); R. Lee *et al.*, *Cell* **75**, 843 (1993).
43. R. Johnson and C. Tabin, *Cell* **81**, 313 (1995); T. R. Burglin, *Curr. Biol.* **6**, 1047 (1996). The HOG genes identified by Burglin bear the probable intein-like autoproteolysis domain that is also present in hedgehog and may also recognize and bind to sterols, but these genes do not bear the many other features of vertebrate and invertebrate hedgehog orthologs.
44. P. Kuwabara, personal communication.
45. B. J Meyer, in C. elegans *II*, D. L. Riddle *et al.*, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1997), pp. 209–240.
46. P. G. Okkema and J. Kimble, *EMBO J.* **10**, 171 (1991).
47. T. M. Barnes and J. Hodgkin, *ibid.* **15**, 4477 (1996); P. Aza-Blanc *et al.*, *Cell* **89**, 1043 (1997).
48. Y. T. Ip and M. Levine, *Curr. Opin. Genet. Dev.* **14**, 672 (1994).
49. X. S. Hou and N. Perrimon, *Trends Genet.* **13**, 105 (1997).
50. A. Fire *et al.*, *Nature* **391**, 806 (1998).
51. G. Jansen *et al.*, *Nature Genet.* **17**, 119 (1997).
52. We have in general referenced reviews and apologize to all those whose work we could not cite due to space constraints. We are indebted to many *C. elegans* and *Drosophila* developmental geneticists for comments on the manuscript and for communicating unpublished results.

Caenorhabditis elegans Is a Nematode

Mark Blaxter

REVIEW

***Caenorhabditis elegans* is a rhabditid nematode. What relevance does this have for the interpretation of the complete genome sequence, and how will it affect the exploitation of the sequence for scientific and social ends? Nematodes are only distantly related to humans and other animal groups; will this limit the universality of the *C. elegans* story? Many nematodes are parasites; can knowledge of the *C. elegans* sequence aid in the prevention and treatment of disease?**

In terms of numbers of described species, the arthropods dominate the known metazoan life on Earth. Although the number of described species of nematode is only ~20,000, estimates of the actual number range from 40,000 to 10 million. The high estimates are based on repeated sampling of single marine habitats and are supported by surveys of terrestrial faunas (*1*). Nematodes are also numerically abundant, attaining millions of individuals per square meter (*2*). *Caenorhabditis elegans* is therefore a representative of a diverse and successful group of animals.

How do the molecular, physiological, and developmental mechanisms used by *C. elegans*—as revealed by the *C. elegans* genome sequence and by the equally important genetic and developmental biological work carried out in the last 30 years (*3*)—relate to those used by other animals? Although there are undoubtedly nematode-specific components to the *C. elegans* basic body plan, some recent studies indicate that signaling systems have been recruited wholesale to perform new functions as if they are self-contained cassettes that can be exchanged with little functional consequence (*4*). At a higher level, though, the patterns and processes used by *C. elegans* to build its body are a product of adaptive evolution over millions of years. Thus, the phylogenetic position of *C. elegans* with respect to other animals is of importance in deciphering the modes and tempos of evolution of these processes (*5*).

For example, if a gene [such as a particular nuclear hormone receptor subtype (*4*)] is found in both the fruit fly *Drosophila* and *C. elegans*, does this imply that it will most likely also be present in the human genome? If *C. elegans*' ancestor diverged before the vertebrate-arthropod split, the answer will be yes. If, as has been suggested, nematodes are more closely related to arthropods than to vertebrates (see below), similarities between *Drosophila* and *C. elegans* may merely reflect their common ancestry. Is *C. elegans* representative of a primitive metazoan, or is it a highly derived organism?

C. elegans' Place in the Tree of Life

The application of the *C. elegans* project to the understanding of other animals, and of humans in particular, is compromised by the deep phyletic separation of the nematodes from other groups. Current best estimates of the time of divergence range from 1200 million to 600 million years ago (*6*). There are about 35 animal groups whose body plans are distinct enough to warrant elevation to phylum status (*7*). After 130 years of phylogeny (*8*), the interrelationships of the animal phyla are still the subject of vigorous debate, and the position of the Nematoda within the animals is far from clear. The integration of molecular and morphological analyses is required to resolve these long-standing problems (*9*).

Morphological phylogenies have usually indicated that the pseudocoelomate nematodes arose early in animal evolution, as part of a radiation of "aschelminth" phyla, predating the split into protostome groups (annelids, arthropods, mollusks, and others) and deuterostome groups (chordates, brachiopods, and others) (Fig. 1A) (*10, 11*). This scheme suggests that nematodes are equally distant from both arthropods and vertebrates. Cladistic analyses of developmental and morphological traits have resulted in a reassessment of this unresolved phylogeny. Nielsen (*7*) proposed that the nematodes, along with four other pseudocoelomate phyla (nematomorphs, priapulids, kinorhynchs, and loriciferans), form a monophyletic group of animals with an introvert (extensible, spined anterior organ), no locomotory cilia, and a cuticle that is shed at periodic molts. Nematodes are recognized as protostomes, animals where the mouth is formed from the embryonic blastopore. This feature is not particularly evident in *C. elegans*, where the embryo is a dense mass of cells and the blastopore is not distinct, but is in other nematodes (*12*). In Nielsen's phylogeny, therefore, nematodes are slightly more closely related to arthropods than they are to vertebrates.

Molecular phylogenetic analyses of the position of the Nematoda with respect to other phyla were initially compromised by the use of *C. elegans* as a marker nematode taxon. The genes of *C. elegans* appear to have undergone accelerated molecular evolution relative to those of many other animals. This relative rate difference resulted in the (probably) artifactual placement of the origin of *C. elegans* (and with it, by association, all of the nematodes) very early in metazoan molecular phylogenies. This phenomenon has meant that the nematodes have been left out of such analyses until recently. Sequencing of small subunit ribosomal RNA genes from additional species of nematode has yielded taxa with reduced apparent rates, and these sequences can be used to place nematodes more robustly within the metazoa (*13, 14*). The results of these studies are surprising and challenge the view that nematodes branched off before the arthropod-vertebrate split. Two major rearrangements are proposed. The arthropods are removed from a close relationship to the annelids, and a new high-level taxon, of animals that shed a cuticle by ecdysis (the Ecdysozoa), is proposed to include arthropods, nematodes, and their allies (Fig. 1C) (*14*). The Ecdysozoa hypothesis is not universally accepted, as it

The author is at the Institute of Cell, Animal, and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK.