Opinion

What does a worm want with 20,000 genes?

Jonathan Hodgkin

Address: Genetics Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK.
E-mail: jah@bioch.ox.ac.uk

Published: 17 October 2001

Genome ***Biology*** 2001, **2(11)**:comment2008.1–2008.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/11/comment/2008

Abstract

The number of genes predicted for the *Caenorhabditis elegans* genome is remarkably high: approximately 20,000, if both protein-coding and RNA-coding genes are counted. This article discusses possible explanations for such a high value.

One of the surprises to be announced at this year's milestone in human genome analysis [1] was the low number predicted for the total number of human genes: somewhere around 30,000. Most earlier estimates had been for a much higher number, in the range 60,000 to 150,000. In contrast, three years ago, when the genome sequence of the nematode *Caenorhabditis elegans* was essentially completed [2], one of the surprises was the high number of predicted genes: approximately 19,000. Most earlier estimates had been much lower: under 10,000. So why do worms have so many genes? Does this mean anything? And is the current number real?

Exactly what the total gene count is for the human genome will probably remain vague for some time to come, but it seems likely that the number will move upwards as the sequence is refined and annotation improves. Numbers are firmer for the *C. elegans* genome, and have not changed much over the past two years. About half of all *C. elegans* genes are currently enigmatic in terms of sequence similarity and function, however, and their existence was initially based only on GeneFinder predictions [2]. According to some analyses, many could be pseudogenes [3].

The question of how many of these predicted worm genes are real has recently been tackled by Reboul *et al.* [4], using an OST (open-reading-frame sequence tag) approach. They chose 1,222 predicted genes for which no EST (expressed sequence tag) had yet been obtained, and attempted to amplify a predicted product from cDNA. At least 70% of the genes were thereby verified, indicating that they are real, although the predicted intron/exon structure was not always correct. Their study resulted in a minimum, and therefore conservative, estimate for the *C. elegans* gene number of 17,387. This can be compared with the most recent genome-sequence-based number of 19,404 protein-coding genes [5], which is not too dissimilar. Both approaches are likely to have missed a substantial number of small genes, such as those encoding neuropeptides, antimicrobial peptides, cuticle components and small regulatory proteins such as *egl-1* [6]. The high total gene count therefore does not seem to be an artifact of the prediction programs, nor to be explained by the presence of numerous pseudogenes.

Moreover, the numbers above apply only to protein-coding genes, and there are a substantial number of RNA-encoding genes which have to be added to the gene tally. These RNA-encoding genes are often difficult to recognize on the basis of sequence, so they are usually even harder to count than the protein-coding genes. Some of the classes, such as ribosomal RNA genes, are easy to enumerate. Worms have one cluster of 55 large ribosomal genes, encoding the 18S, 5.8S and 28S rRNAs, and another cluster of 110 genes for 5S rRNA. Transfer RNA genes are for the most part also fairly easy to recognize, and have been extensively annotated in the worm genome. About 900 tRNA genes can be recognized; 200 of these are probably pseudogenes, but the majority look real. This is consistent with the one family that has been examined in detail (the twelve Trp tRNA genes), in which eight are demonstrably functional, and two look like pseudogenes [7]. The relative number of tRNA genes is high compared to the number in the human genome, in one of the many small mysteries of comparative genomics.

Other RNA genes include snRNAs (small nuclear RNAs), snoRNAs (small nucleolar RNAs), scRNAs (small cytoplasmic RNAs), telomere RNAs, splice leaders and small regulatory RNAs. Some of these, such as the snoRNAs, are hard to recognize in raw sequence [8], and consequently their exact number is unknown. Others, such as the developmental timing regulators *lin-4* and *let-7,* are known to perform important biological functions, but their discovery has depended entirely on genetic methods [9]. It is an open question how many other non-coding regulatory RNAs like these remain to be found in eukaryotic genomes. The two known examples of regulatory RNAs in *C. elegans* represent fewer than 0.5% of the genes defined by mutation and subsequently cloned, which implies that regulatory RNAs cannot be all that numerous, but that still leaves room for many more genes in this category. In summary, the number of known RNA genes is already well over 1,000. These can be added to the 18,000-19,000 predicted protein-coding genes, to give a total of something like 20,000 as a nice round number - hence the title of this article.

What are all these genes doing? This question is being attacked systematically by several kinds of functional test. Ideally, clean deletional knockouts of every gene in the worm would be made, comparable to the set that has been created for the budding yeast *Saccharomyces cerevisiae* [10], but efficient gene disruption by homologous recombination is not yet feasible in *C. elegans*. Consequently, systematic deletion is more work and chancier in worms than in yeast and progress is slower. Nevertheless, several hundred gene deletions have been generated already, and improvements in technology are leading to further increases in the knockout production rate [11].

While the research community waits for that resource, we are already provided with a cornucopia of information arising from expression studies [12], SAGE analysis [13], microarray data [14] and especially from the use of RNAi to generate transient knockouts [15]. RNAi knockouts entail treating worms with double-stranded (ds) RNA corresponding to coding sequences in an endogenous gene. This usually leads to a massive reduction in expression from the target gene, apparently by selective degradation of its mRNA. In high-throughput RNAi experiments, the dsRNA has been applied by microinjection [16], by feeding the worms on bacteria expressing a particular dsRNA [17], or by soaking worms directly in dsRNA solutions [18]. The technique is known to be fallible, in that some genes are refractory to RNAi (notably those expressed in neurons), but it probably works on well over 50% of all worm genes. Nevertheless, fewer than 15% of targets tested by RNAi yield any obvious phenotype. Similar conclusions emerge from the deletional knockout program. Further confirmation is provided by an increasing number of examples of homozygous-viable mutations in known genes, which have turned out to be small deletions removing several adjacent genes, without any significant additional phenotypes.

In sum, all of these studies indicate that most worm gene knockouts result in no obvious change in development, viability or behavior. This is consistent with the previous expectations from classical mutational analyses of *C. elegans,* which indicated that its genome contained fewer than 5,000 singly essential genes, and fewer than 7,000 genes for which mutational loss would have a noticeable effect. What, then, are the explanations for a total gene number three times higher? Some possible factors are discussed below, any or all of which may contribute to the high gene number.

One factor frequently suggested is a relative lack of alternative splicing in *C. elegans* compared with *Drosophila* or mammals. The generation of more than one polypeptide from a given primary RNA transcript by means of alternative splicing provides a means of greatly amplifying the number of different proteins in an animal, and it is guessed that the 30,000-40,000 genes in mammals may generate a total repertoire of 100,000 or more different final proteins. In extreme cases, it may be that a single gene can generate more than 1,000 different isoforms by means of alternative splicing. In *C. elegans,* the phenomenon certainly occurs, but probably to a lesser degree than in mammals. Confirmed cases of alternative splicing amount to a current total of about 4% of genes (815/19,404) in the worm [5]. Detection of different isoforms is largely dependent on cDNA data, however, and it is already apparent that the current EST databases for *C. elegans* are not adequately representative, because so far at least 40% of the genes in the organism are not reflected in these databases. Detection of significant isoforms for some genes will be particularly difficult if they are present only at low levels or only in a few cells. Current prediction programs are notoriously bad at predicting or evaluating alternative splicing in any multicellular organism. For these reasons, it is difficult to estimate the percentage of genes experiencing alternative splicing in the nematode: all one can say is that it is certainly higher than the current value, and probably lower than in mammals. This 'splicing factor' provides a reasonable explanation for why mammals have unexpectedly few genes, but not for why nematodes have unexpectedly many.

A second, and probably related, effect is that proteins encoded by the human genome tend to contain more multiple domains than do those encoded by *C. elegans*. This is especially true for transmembrane proteins with large extracellular regions, which are presumably involved in cell-cell interactions of one kind or another. Such classes of nematode proteins may be generally simpler and less multifunctional than their mammalian counterparts, and therefore a greater variety would need to be encoded by separate genes.

A third consideration arises from the apparent extensive duplication of genes in the *C. elegans* genome. Many genes and small genetic regions appear to have experienced piecemeal duplication during the evolution of this species [2]. The

level of duplication appears to be higher in the autosomal chromosome-arm regions of the genome, which are also regions where genes encoding conserved eukaryotic core functions are underrepresented [2,19]. Much of the duplication seems to be of ancient origin, which implies that selective forces maintain the present high number of gene pairs, whatever their origins may be.

The question of functional redundancy in the duplicated genes and regions becomes relevant here. Most of the current examples of overlap in function between members of gene families in *C. elegans* conform to a pattern of incomplete redundancy. That is, two genes may overlap in some of their functions, but each has at least one unique function. A good example is provided by the two Notch-related receptor genes of the worm, *lin-12* and *glp-1,* which have distinct postembryonic roles but share some functions during embryogenesis [20]. Such an arrangement is both evolutionarily stable and advantageous, because neither gene can drift into dysfunctionality, while the shared functions are made more robust. Whether and why this may have happened more frequently in nematodes than in other kinds of animal is not clear.

Redundant genetic programming, and a concomitant high gene number, may underlie much of the conspicuous invariance observed in the development and behavior of nematodes. These organisms may in fact be subject to greater developmental constraints than other animal phyla, and it is conceivable that these constraints may in turn have resulted in selection for further increases in gene number, especially for neuronal genes. Strikingly, the nervous system of all nematode species (which may exceed 10 million in total [21]) contain remarkably few neurons, and there is conspicuous conservation of neuronal anatomy and connectivity between the parasitic worm *Ascaris suum* and *C. elegans,* despite the large differences in size and behavior between these two well-studied species. Functional analysis of *C. elegans* neurons suggests that many of these neurons are multitasked, so that a single cell may be responsible for more than one piece of behavior. The most impressive example of multi-tasking is provided by the olfactory repertoire: the number of different worm chemoreceptor genes greatly exceeds the number of olfactory neurons, each of which must therefore express many different receptors [22]. Yet the worm is able to distinguish different odors quite efficiently, presumably by using sophisticated signal transduction machinery. Evolution may therefore have adopted a pattern of increased biochemical complexity to compensate for an inherent lack of anatomical complexity in the nematode nervous system.

The olfactory receptors represent, collectively, the most inflated set of *C. elegans* genes, and there are indications that the frequency of pseudogenes is higher among them than in other gene sets [23]. This observation raises the possibility that there has been past selection for a considerable expansion - an efflorescence - of these receptor genes, but that selective forces are no longer maintaining such a high number. The same could be true for some of the other conspicuously large gene families in *C. elegans,* such as the *nhr* genes (encoding nuclear hormone receptors), which are much more numerous in *C. elegans* than in either the fly or human genomes [24]. A plausible example of family expansion is provided by the 150-odd *msp* genes that encode the major sperm protein of *C. elegans*, many of which are pseudogenes. In contrast, this protein is encoded by only two or three genes in *Ascaris* [25]. The expansion has probably resulted from the evolution of a self-fertilizing hermaphrodite sex in *C. elegans,* for which rapid production of sperm is essential. Transient efflorescence of different gene families might be an important factor in increasing the gene number of the worm.

Important new information on the scale of most of the above factors (possible pseudogenes, alternative splicing, duplication, redundancy and efflorescence) will soon emerge from near-complete sequencing of a close relative of *C. elegans,* the nematode *Caenorhabditis briggsae*. Before this year, more than 10% of the *C. briggsae* genomic sequence was already available (Genome Sequencing Center, Washington University, unpublished data). By the end of 2001, 10x 'shotgun' sequence coverage of the whole *C. briggsae* genome should be finished, and these 'random' pieces of sequence may allow the assembly, through overlaps, of a genomic sequence almost as complete as that of *C. elegans*. The two species are extremely similar in development and morphology (though incapable of interbreeding, alas), yet have diverged by 20 million years or more of separate evolution, so the patterns of conservation and difference between the two nematode genomes will be hugely informative.

A final factor that may contribute to the high gene number of *C. elegans* is its ecology, which is something we know remarkably little about. The preferred ecological niche for this soil-dwelling species seems to be the exploitation of brief bacterial blooms associated with decaying vegetable matter, especially mushrooms. The extreme rapidity of the growth of *C. elegans,* which makes it so beloved of geneticists and developmental biologists, probably results from specialization for this niche. But as a soil organism, *C. elegans* has to survive in an immensely complex ecosystem, and moreover one that is going to differ considerably from one geographical location to another. *C. elegans* has been recovered from locations all over the world, so it evidently has the ability to deal with many different climates [26]. As a bacterial feeder, it will be constantly challenged by all the different species of soil bacteria, fungi and other microbes. This is a far cry from the monotonous and unnatural diet of *Escherichia coli* that laboratory strains get to eat. The use of *C. elegans* to investigate its interactions with other bacteria, some of them medically or agriculturally important, is a recent and exciting development in the field of *C. elegans*

biology [27]. Tan *et al*. [28] have demonstrated the potential for investigating bacterial pathogenicity by exposing worms to the medically important *Pseudomonas aeruginosa* PA14, and Marroquin *et al*. [29] have shown that *C. elegans* can be used to study mechanisms of resistance to Bt toxin from *Bacillus thuringiensis*, a hazard that undoubtedly exists in the natural environment of the worm. At least five loci can mutate to confer resistance to one of the Bt toxins but the resistant mutants appear normal in other respects and would have been missed in any standard screen for visible mutants [29]. A similar story is emerging from investigations of interactions between *C. elegans* and a specific bacterial pathogen for this species, *Microbacterium nematophilum* [30]. Mutants resistant to infection by *M. nematophilum* define at least 20 genes, and many of these mutants exhibit no other conspicuous phenotype (M. Gravato-Nobre and J.H., unpublished observations). The recent accidental discovery of *M. nematophilum* raises the possibility that many other unidentified nematode pathogens exist in the soil. Any or all of these pathogens may be driving the evolution of *C. elegans,* in ways that will only be comprehensible with better knowledge of its natural environment.

An ironic extension of this kind of thinking is that there is only one organism that we have any hope of understanding fully, in terms of biology. That organism is our own species. Only we are in a position to report on every disease or toxin we encounter. Similarly, only we can adequately monitor our own physiological and genomic responses to arctic blizzards, tropical heat, psychological stresses and social pleasures. And only for human beings will we ever have truly extensive data on genomic variability and molecular paleontology. The other tens of millions of species on this planet are likely to remain largely mysterious forever, especially as most of them may well go extinct during the present century, along with all their complex biotic interactions. The biological world will then become more comprehensible but incomparably poorer.

References

1. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
2. The *C. elegans* Sequencing Consortium: **Genome sequence for the nematode *C. elegans:* a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
3. Harrison PM, Echols N, Gerstein MB: **Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome.** *Nucleic Acids Res* 2001, **29:**818-830.
4. Reboul J, Vaglio P, Tzellas N, Thierry-Mieg N, Moore T, Jackson C, Shin-i T, Kohara Y, Thierry-Mieg D, Thierry-Mieg J, *et al.*: **Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans.*** *Nat Genet* 2001, **27**:332-336.
5. **WormBase** [http://www.wormbase.org/]
6. Conradt B, Horvitz HR: ***C. elegans* protein EGL-1 is required for programmed cell death and interacts with the Bcl-2-like protein CED-9.** *Cell* 1998, **93:**519-529.
7. Eddy SR: **Noncoding RNA genes.** *Curr Opin Genet Dev* 1999, **9:**695-699.
8. Kondo K, Makovec B, Waterston RH, Hodgkin J: **Genetic and molecular analysis of eight tRNA(Trp) amber suppressors in *Caenorhabditis elegans*.** *J Mol Biol* 1990, **215:**7-19.
9. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans.*** *Nature* 2000, **403:**901-906.
10. ***Saccharomyces* Genome Database** [http://genome-www.stanford.edu/Saccharomyces]
11. Liu LX, Spoerke JM, Mulligan EL, Chen J, Reardon B, Westlund B, Sun L, Abel K, Armstrong B, Hardiman G, *et al.*: **High-throughput isolation of *Caenorhabditis elegans* deletion mutants.** *Genome Res* 1999, **9:**859-867.
12. Hope IA, Albertson DG, Martinelli SD, Lynch AS, Sonnhammer E, Durbin R: **The *C. elegans* expression pattern database: a beginning.** *Trends Genet* 1996, **12:**370-371.
13. Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA: **Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans.*** *Genome Res* 2001, **11:**1346-1352.
14. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: **A gene expression map for *Caenorhabditis elegans.*** *Science* 2001, **293:**2087-2092.
15. Fire A**: RNA-triggered gene silencing.** *Trends Genet* 1999, **15:**358-363.
16. Gonczy P, Echeverri G, Oegema K, Coulson A, Jones SJ, Copley RR, Duperon J, Oegema J, Brehm M, Cassin E, *et al.*: **Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III.** *Nature* 2000, **408:**331-336.
17. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J: **Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference.** *Nature* 2000, **408:**325-330.
18. Maeda I, Kohara Y, Yamamoto M, Sugimoto A: **Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi.** *Curr Biol* 2001, **11:**171-176.
19. Hutter H, Vogel BE, Plenefisch JD, Norris CR, Proenca RB, Spieth J, Guo C, Mastwal S, Zhu X, Scheel J, Hedgecock EM: **Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes.** *Science* 2000, **287:**989-994.
20. Kimble J, Simpson P: **The LIN-12/Notch signaling pathway and its regulation.** *Annu Rev Cell Dev Biol* 1997, **13:**333-361.
21. Blaxter M: ***Caenorhabditis elegans* is a nematode.** *Science* 1998, **282:**2041-3046.
22. Bargmann CI: **Neurobiology of the *Caenorhabditis elegans* genome.** *Science* 1998, **282:**2028-2033.
23. Robertson HM: **The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses.** *Genome Res* 2000, **10:**192-203.
24. Maglich JM, Sluder A, Guan X, Shi Y, McKee DD, Carrick K, Kamdar K, Willson TM, Moore JT: **Comparison of complete nuclear receptor sets from the human, *Caenorhabditis elegans* and *Drosophila* genomes.** *Genome Biol* 2001 **2**: research0029.1-0029.7.
25. Bennett KL, Ward S: **Neither a germ line-specific nor several somatically expressed genes are lost or rearranged during embryonic chromatin diminution in the nematode *Ascaris lumbricoides* var. *suum*.** *Dev Biol* 1986, **118:**141-147.
26. Hodgkin J, Doniach T: **Natural variation and copulatory plug formation in *Caenorhabditis elegans.*** *Genetics* 1997, **146:**149-64.
27. Kurz CL, Ewbank JJ: ***Caenorhabditis elegans* for the study of host-pathogen interactions.** *Trends Microbiol* 2000, **8:**142-144.
28. Tan MW, Rahme LG, Sternberg JA, Tompkins RG, Ausubel FM: ***Pseudomonas aeruginosa* killing of *Caenorhabditis elegans* used to identify *P. aeruginosa* virulence factors.** *Proc Natl Acad Sci USA* 1999, **96:**2408-2413.
29. Marroquin LD, Elyassnia D, Griffitts JS, Feitelson JS, Aroian RV: ***Bacillus thuringiensis* (Bt) toxin susceptibility and isolation of resistance mutants in the nematode *Caenorhabditis elegans*.** *Genetics* 2000 **155:**1693-1699.
30. Hodgkin J, Kuwabara PE, Corneliussen B: **A novel bacterial pathogen, *Microbacterium nematophilum*, induces morphological change in the nematode *C. elegans.*** *Curr Biol* 2000, **10:**1615-1618.