The truth about mouse, human, worms and yeast

*David R. Nelson[1] and Daniel W. Nebert[2]**

[1]Department of Molecular Sciences and The UT Center of Excellence in Genomics and Bioinformatics, University of Tennessee, Memphis, Tennessee 38163, USA
[2]Department of Environmental Health and Center for Environmental Genetics (CEG), University of Cincinnati Medical Center, Cincinnati, Ohio 45267-0056, USA
**Correspondence to:* Tel: +1 513 558 4347; Fax: +1 513 558 3562; E-mail: dan.nebert@uc.edu

Date received: 1st December 2003

Abstract

Genome comparisons are behind the powerful new annotation methods being developed to find all human genes, as well as genes from other genomes. Genomes are now frequently being studied in pairs to provide cross-comparison datasets. This 'Noah's Ark' approach often reveals unsuspected genes and may support the deletion of false-positive predictions. Joining mouse and human as the cross-comparison dataset for the first two mammals are: two *Drosophila* species, *D. melanogaster* and *D. pseudoobscura*; two sea squirts, *Ciona intestinalis* and *Ciona savignyi*; four yeast (*Saccharomyces*) species; two nematodes, *Caenorhabditis elegans* and *Caenorhabditis briggsae*; and two pufferfish (*Takefugu rubripes* and *Tetraodon nigroviridis*). Even genomes like yeast and *C. elegans*, which have been known for more than five years, are now being significantly improved. Methods developed for yeast or nematodes will now be applied to mouse and human, and soon to additional mammals such as rat and dog, to identify all the mammalian protein-coding genes. Current large disparities between human Unigene predictions (127,835 genes) and gene-scanning methods (45,000 genes) still need to be resolved. This will be the challenge during the next few years.

Keywords: *human genome, mouse genome, Caenorhabditis elegans genome, Caenorhabditis briggsae genome, Saccharomyces genomes, comparative genomics, gene discovery, gene-prediction algorithms*

Introduction and background

The monumental sequence of a composite human genome conjures up images of Arthur C. Clark's monolith in the film *2001: A Space Odyssey* — a beautiful, awe-inspiring structure with a hidden message. Researchers' ignorance is laid bare by the simple fact that they cannot, with any confidence, extract from this (human genome) structure the total number of its genes. What is needed is a Carl Sagan (*SETI, Contact*), an Alan Turing (WWII code breaker) or a Jean-François Champollion (Rosetta stone decoder) to break the codes. Or perhaps, what is really needed is a Rosetta stone for genomes: just two or three translations of the same message, laid side by side. Unfortunately, there is not even one full translation available. James Watson put it this way in a 1992 interview:[1] 'The goal of the Human Genome Project is to understand the genetic instructions for human beings . . . Getting the instructions is a big job; understanding those instructions can consume many hundreds of years . . .'.

In December 1999, an analysis of the human chromosome (Chr) 22 sequence was published; 545 protein-coding genes and 134 pseudogenes were identified.[2] In January 2003, a reanalysis of the Chr 22 sequence by the same group reported 546 protein-coding genes and 234 pseudogenes, with an increase of 74 per cent in the total length of exons in the annotation.[3] A third, microarray-based, study[4] doubled the number of Chr 22 base pairs in transcribed sequences. The National Center for Biotechnology Information (NCBI) human genome map-viewer build 34 version 1 (Nov 2003) has 673 genes on Chr 22 and an unspecified number of pseudogenes. Since the true number of genes and pseudogenes has not changed in the past four years — it is merely researchers' ability to detect them that has improved — how many more genes will be found and how will they be detected?

Finding protein-coding genes

The best method for documenting genes is with a full-length cDNA. Even shorter expressed-sequence tags (ESTs), if not from the same species then from a closely related species, are useful. The EST database dbEST (21st November, 2003) lists 5,427,257 *Homo sapiens* ESTs and 3,948,029 *Mus musculus*

ESTs. The Unigene database clusters these ESTs into unique contigs representing 127,835 human (build 163) and 93,645 mouse transcripts. The human number is similar to the TIGR Gene Index prediction of 120,000 genes in humans.[5]

According to the NCBI Handbook 2003, Unigene clusters may contain more than one alternative-splice form.[6] Furthermore, Unigene clusters are required to have evidence of a 3′ terminus, to avoid forming two or more clusters from a single long gene; this restriction prevents some ESTs in dbEST from being included in Unigene. The logical interpretation[7] of these facts is that 'each Unigene cluster contains sequences that represent a unique gene'.

This leaves researchers with a problem. Conservative gene annotation of the human genome only identified 25,642 genes.[8] More relaxed estimates predict about 40,000[9] to 45,000[10] genes; yet, these numbers are about threefold lower than the Unigene cluster count. At some point, these values should converge on the true number of genes — defined as full-length, expressed messages from any cell type at any time, from germ cells to embryo to adult. Currently, this point is some way away.

By the comparative genomics approach, the mouse genome is supposed to save us from this weakness in finding genes in the human genome. By comparing mouse and human genomic sequences, all orthologous genes and many paralogous genes should be detectable, exon by exon. Preliminary efforts with small sets of known genes were highly successful. The ROSETTA program[11] identified 94 per cent of internal coding exons from 117 mouse–human orthologous gene pairs perfectly at both exon ends, and another 4 per cent at one of the two ends.[12] It did less well for initial, plus terminal, coding exons.

Including conserved sequence elements

We now find the problem grows more complex, however, because there are thousands of non-expressed conserved sequence elements (CSEs) in the two mammals,[9] sequences whose function we do not understand. Some are possibly promoter regions, some pseudogenes or RNA genes and some are new undocumented genes, but it is clear that this does not account for all of these sequences. Thus, the comparative genomics approach may over-predict, when viewing two mammals, since they may be phylogenetically too close. The distance between species for optimal gene identification has been studied, and mouse–human is generally good, but a mammal more distant than mouse from human might be even better.[13]

An alternative approach has been to use fish as a more distant relative. The EXOFISH Program[14] compared human and *Tetraodon nigroviridis* (freshwater pufferfish) for conserved regions (presumably exons) and found 28,000–34,000 genes. Due to the greater evolutionary distance between human and fish, there is a cleaner background, but the many mammal-specific genes and human brain-specific genes may not be identified, so the gene number predicted by EXOFISH is almost certainly an underestimate.

Another approach is exemplified by the analysis of sequences from 12 species, all derived from a 1.8 Mb region orthologous to a human Chr 7 segment containing ten genes.[15] In this instance, coding exons were already well documented, but substantial numbers of CSEs — beyond those previously identified experimentally — were discovered. This approach might be more fruitful at human gene discovery, if applied to areas of the human genome that are more poorly characterised than the Chr 7 segment chosen.

Whereas ~1.5 per cent of the human genome comprises protein-coding genes, another ~3.5 per cent of the genome contains CSEs that are more conserved than protein-coding-gene regions.[16] Possible functions for these CSEs (termed CNGs by Dermitzakis *et al.*[16] and CNSs by Inada *et al.*[17]) include control regions that: (a) regulate gene expression; (b) govern developmental-, cell type- and organ-specific expression, in trans, of genes located far away; (c) lock-in regulatory decisions;[17] and (d) act as structural components of chromosomes when alignment and chromosome movement occurs during meiosis or mitosis. There appear to be at least twice as many CSEs than protein-coding genes in the genome. A recent comparison of 43 species — including vertebrates, insects, worms, plants, fungi, yeast, eubacteria and archaebacteria[18] — revealed noteworthy increases in genome size and complexity from prokaryote to mammals, again emphasising the innumerable highly-conserved CSEs that are likely to have essential functions and critical effects on an organism's phenotype.

Learning from the worms

Nevertheless, the comparative approach is very powerful — as illustrated by the recent comparative genomics study of *Caenorhabditis elegans* and *Caenorhabditis briggsae*.[19] This study increased the signal-to-noise ratio by using four gene-prediction algorithms on each genome, comparing results between genomes and selecting the most informative dataset (Figure 1). The power of this method was proven by the prediction of 1,275 new genes in *C. elegans* that had not been detected in the previous five years of annotation on this genome (Table 1). Huge numbers of previously predicted *C. elegans* genes were also revised, due to the identification of many new exons, based on these novel findings.[19]

Different algorithms for predicting protein-coding genes give similar results in predicting exons but tend to disagree on the grouping of exons into genes.[20] Four different gene-prediction programs can give four very different answers across the same region of a genome. Stein *et al.*[19] used the concordance of prediction between *C. elegans* and *C. briggsae* to

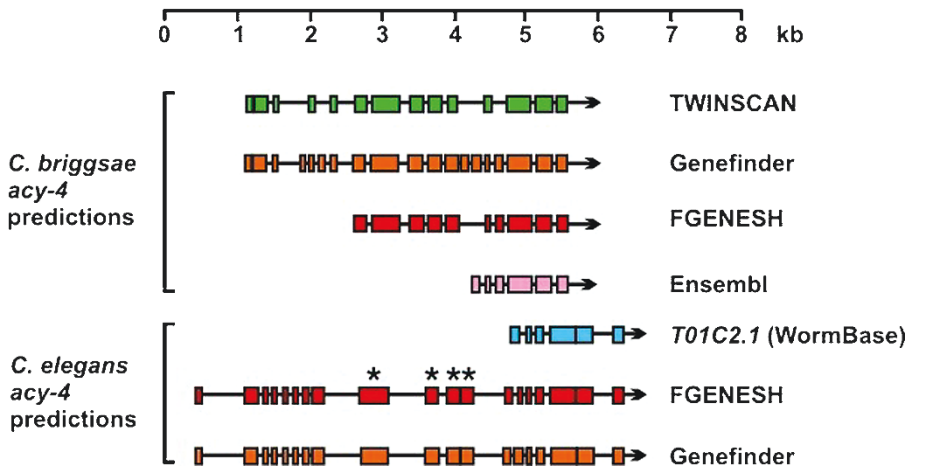

Figure 1. Use of the four gene-prediction algorithms to characterise the *acy-4* gene in both worm genomes. Of the 12 possible combinations of predictions, the Genefinder-Genefinder prediction pair was chosen as the best model, because this pair showed the greatest similarity to each other, excluding terminal exons. Coding sequence conservation between the two has provided evidence for as many as 12 additional N-terminal exons in the Genefinder *Caenorhabditis elegans acy-4* prediction, compared with that of *T01C2.1*, the WormBase WS77 *C. elegans acy-4* prediction. Four of the additional N-terminal exons (those marked with asterisks) that were predicted by FGENESH and Genefinder have subsequently been confirmed by new EST data (*modified from* Ref. [19])

predict the most likely gene model — using Genefinder (*version 980506*, P. Green, unpublished data, 2003; see also Ref. [7]), FGENESH,[21] TWINSCAN[22] and the Ensembl annotation pipeline.[23] The output of the four gene-prediction programs (Figure 1) was largely concordant with respect to the position of *C. briggsae* exons (80 per cent of exons predicted identically by two or more programs; 26 per cent predicted identically by all four programs), but discordant with regard to gene predictions (38 per cent of genes called identically by two or more programs; just 4 per cent called identically by all four programs). A similar pattern was seen in *C. elegans*.[19]

Stein *et al.*[19] termed the gene sets produced by their analysis 'hybrid gene sets', because the final gene sets are a mixture of gene prediction from multiple programs; applying a transposon- and pseudogene-filtering step to the WormBase 77 set, they removed 619 genes to create a 'pruned' WS77 set, termed WS77⋆. The constitution of the final gene sets was: *C. briggsae*, 19,507 genes; the *C. elegans* WS77⋆, 18,808 genes; and the hybrid *C. elegans*, 20,621 genes (Table 1).

Stein *et al.*[19] compared the *C. elegans* hybrid gene set (20,621 genes) to the WS77⋆ set (18,808 genes) derived from WormBase and derived 1,275 well-supported suggestions for new *C. elegans* genes, 1,763 new exons in 1,100 existing genes, 2,093 exon deletions in 1,583 genes, 1,675 exon truncations in 1,502 existing genes and 1,115 exon extensions in 1,008 existing genes. These data underscore the value of comparative genomics between total-genome sequences from two species in establishing a more accurate count of protein-coding genes.

Comparing *C. elegans*/*C. briggsae* divergence and mouse/human divergence

The two worms diverged ~100 million years ago (MYA) and the two mammals diverged ~75 MYA. Similar levels of amino acid identity exist between *C. briggsae* and *C. elegans* orthologues (80 per cent) and between mouse and human orthologues (78.5 per cent). In the mouse/human comparison, 80 per cent of predicted proteins can be assigned to a 1:1 orthologue pair, whereas <65 per cent of *C. briggsae* genes could be assigned a *C. elegans* orthologue. The protein families

Table 1. Several comparisons of *Caenorhabditis briggsae*, *Caenorhabditis elegans* WS77* and *Caenorhabditis elegans* hybrid

Category	*C. briggsae*	*C. elegans* WS77*	*C. elegans* hybrid
Genome size	~104 Mb	100.3 Mb	–
Number of genes	19,507	18,808	20,621
Number of exons	114,339	118,045	125,702

Data taken from Stein *et al.*[19]

are thus more dynamic in the two nematodes — several hundred either being novel or having diverged so far that their common origin cannot be recognised, and another ~200 having expanded or contracted by more than twofold. The *C. briggsae*/*C. elegans* pair is also evolving more rapidly at the nucleotide level: 1.78 synonymous substitutions per synonymous site, compared with 0.6 in the mouse/human pair.[19]

Many of these striking differences between the two worms and the two mammals can probably be explained on the length of generation times. The generation time in the nematodes is ~3 days, compared with ~3 months and ~20 years for the mouse and human, respectively.

Approaching a stable gene count in yeast: Hope for mammals

Improved annotation does not always increase gene number. Detailed comparison of four *Saccharomyces* species[24] resulted in revision of 15 per cent of known yeast genes and a net decrease in the *S. cerevisiae* gene count of about 500; this is a case where 'less is more'. This illustrates the power of adding more closely related sequences to the analysis, especially since the yeast genome had been known for seven years prior to this analysis.

Conclusions

Tremendous progress has been made in the eight years since the baker's yeast genome sequence appeared. There is still a large gap, however, between gene predictions and Unigene clusters. This must be accounted for by improvement of comparative genomics methods such as: (a) using the ROSETTA program to include three or more species; (b) obtaining more comprehensive EST collections from mouse, rat, human and other species, possibly by purchase of these resources from private companies that have already amassed the information; and/or (c) utilising consensus prediction methods, as was done in the *C. elegans*/*C. briggsae* study.[18] Special attention will need to be given to the first- and last-exon predictions, as well as allowance of non-canonical intron–exon boundaries (GC versus GT, etc) — if supported by EST data. Verification of predictions by reverse transcriptase polymerase chain reaction, as was demonstrated in the study by Guigo *et al.*[25] will confirm the expression of questionable genes and enhance genome annotation. One can only hope that Dr Watson's prediction of 12 years ago was a slight exaggeration.

Acknowledgments

The writing of this article was funded, in part, by NIH grant P30 ES06096 (D.W.N.). The authors very much appreciate assistance with the graphics from Dr Marian Miller.

References

1. http://www.accessexcellence.org/AB/CC/watson.html
2. Dunham, I., Shimizu, N., Roe, B.A. *et al.* (1999), 'The DNA sequence of human chromosome 22', *Nature* Vol. 402, pp. 489–495.
3. Collins, J.E., Goward, M.E., Cole, C.G. *et al.* (2003), 'Reevaluating human gene annotation: A second-generation analysis of chromosome 22', *Genome Res.* Vol. 13, pp. 27–36.
4. Rinn, J.L., Euskirchen, G., Bertone, P. *et al.* (2003), 'The transcriptional activity of human chromosome 22', *Genes Dev.* Vol. 17, pp. 529–540.
5. Liang, F., Holt, I., Pertea, G. *et al.* (2000), 'Gene index analysis of the human genome estimates approximately 120,000 genes', *Nat. Genet.* Vol. 25, pp. 239–240.
6. Pontius, J.U., Wagner, L. and Schuler, G.D. (2003), 'UniGene: A unified view of the transcriptome', in *The NCBI Handbook*, National Center for Biotechnology Information, Bethesda, MD, pp. 21–24.
7. Wheeler, D.L., Church, D.M., Federhen, S. *et al.* (2003), 'Database resources of the National Center for Biotechnology', *Nucleic Acids Res.* Vol. 31, pp. 28–33.
8. Flicek, P., Keibler, E., Hu, P. *et al.* (2003), 'Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map', *Genome Res.* Vol. 13, pp. 46–54.
9. Xuan, Z., Wang, J. and Zhang, M.Q. (2003), 'Computational comparison of two mouse draft genomes and the human golden path', *Genome Biol.* Vol. 4, pp. R1.
10. Das, M., Burge, C.B., Park, E. *et al.* (2001), 'Assessment of the total number of human transcription units', *Genomics* Vol. 77, pp. 71–78.
11. http://www.theory.lcs.mit.edu/crossspecies
12. Batzoglou, S., Pachter, L., Mesirov, J.P. *et al.* (2000), 'Human and mouse gene structure: Comparative analysis and application to exon prediction', *Genome Res.* Vol. 10, pp. 950–958.
13. Zhang, L., Pavlovic, V., Cantor, C.R. and Kasif, S. (2003), 'Human-mouse gene identification by comparative evidence integration and evolutionary analysis', *Genome Res.* Vol. 13, pp. 1190–1202.
14. Roest-Crollius, H., Jaillon, O., Bernot, A. *et al.* (2003), 'Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence', *Nat. Genet.* Vol. 25, pp. 235–238.
15. Thomas, J.W., Touchman, J.W., Blakesley, R.W. *et al.* (2003), 'Comparative analyses of multi-species sequences from targeted genomic regions', *Nature* Vol. 424, pp. 788–793.
16. Dermitzakis, E.T., Reymond, A., Scamuffa, N. *et al.* (2003), 'Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs)', *Science* Vol. 302, pp. 1033–1055.
17. Inada, D.C., Bashir, A., Lee, C. *et al.* (2003), 'Conserved noncoding sequences in the grasses', *Genome Res.* Vol. 13, pp. 2030–2041.
18. Lynch, M. and Conery, J.S. (2003), 'The origins of genome complexity', *Science* Vol. 302, pp. 1401–1404.
19. Stein, L.D., Bao, Z., Blasiar, D. *et al.* (2003), 'The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics', *PLoS. Biol.* Vol. 1, p. 2.
20. Reese, M.G., Hartzell, G., Harris, N.L. *et al.* (2000), 'Genome annotation assessment in *Drosophila melanogaster*', *Genome Res.* Vol. 10, pp. 483–501.
21. Salamov, A.A. and Solovyev, V.V. (2000), '*Ab initio* gene finding in *Drosophila* genomic DNA', *Genome Res.* Vol. 10, pp. 516–522.
22. Korf, I., Flicek, P., Duan, D., *et al.* (2001), 'Integrating genomic homology into gene structure prediction', *Bioinformatics* Vol. 17(Suppl. 1), pp. S140–S148.
23. Clamp, M., Andrews, D., Barker, D., *et al.* (2003), 'Ensembl 2002: Accommodating comparative genomics', *Nucleic Acids Res.* Vol. 31, pp. 38–42.
24. Kellis, M., Patterson, N., Endrizzi, M., *et al.* (2003), 'Sequencing and comparison of yeast species to identify genes and regulatory elements', *Nature* Vol. 423, pp. 241–254.
25. Guigo, R., Dermitzakis, E.T., Agarwal, P., *et al.* (2003), 'Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes', *Proc. Natl Acad. Sci. USA* Vol. 100, pp. 1140–1145.