Selection for short introns in highly expressed genes

Cristian I. Castillo-Davis[1], Sergei L. Mekhedov[2], Daniel L. Hartl[1], Eugene V. Koonin[2] & Fyodor A. Kondrashov[2]

Published online: 22 July 2002, doi:10.1038/ng940

Transcription is a slow and expensive process: in eukaryotes, approximately 20 nucleotides can be transcribed per second[1,2] at the expense of at least two ATP molecules per nucleotide[3]. Thus, at least for highly expressed genes, transcription of long introns, which are particularly common in mammals, is costly. Using data on the expression of genes that encode proteins in *Caenorhabditis elegans* and *Homo sapiens*, we show that introns in highly expressed genes are substantially shorter than those in genes that are expressed at low levels. This difference is greater in humans, such that introns are, on average, 14 times shorter in highly expressed genes than in genes with low expression, whereas in *C. elegans* the difference in intron length is only twofold. In contrast, the density of introns in a gene does not strongly depend on the level of gene expression. Thus, natural selection appears to favor short introns in highly expressed genes to minimize the cost of transcription and other molecular processes, such as splicing.

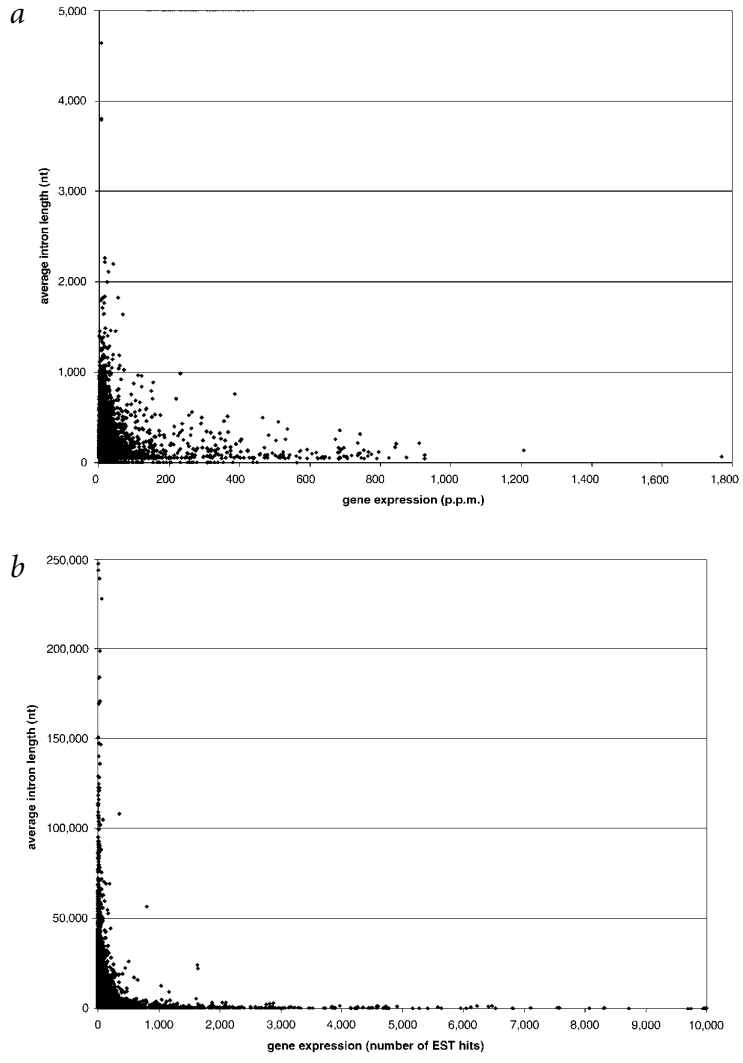

Fig. 1 Scatter plots of the average intron length in a gene versus gene expression. ***a***, Genes from *C. elegans*. ***b***, Genes from *H. sapiens*. p.p.m., parts per million.

Introns are ubiquitous in eukaryotes, although their sizes vary considerably within a genome as well as between different species[4]. Some of the largest introns are found in the human genome, where the total length of intron sequences in a gene often reaches tens of thousands of nucleotides[5] such that transcription of a single gene requires several minutes and thousands of ATP molecules. Mutational preference for deletions may be the sole control on intron size[6,7]. Given the high cost of transcription, however, one might expect that natural selection would favor shorter and fewer introns, especially in genes that, occasionally or constitutively, are expressed at a high level.

To assess gene expression in *C. elegans*, we used data generated by whole-genome analysis, using microarrays[8]. On the basis of the abundance of expressed sequence tags (ESTs) in available EST libraries, we obtained expression data for human genes. We related the expression of genes in the nematode and human to available data on gene structure. Genes expressed at low levels contained a diversity of intron lengths, whereas most highly expressed genes had only short introns (Fig. 1), such that average intron length was significantly shorter in the highly expressed genes (Fig. 2; Table 1). The difference between genes expressed at high and low levels was equally pronounced when we compared the median intron lengths (data not shown). The magnitude of the difference in average intron length between genes expressed at high and low levels is much greater in humans than in nematodes (a factor of 14 compared with a factor of 2), perhaps because the introns in human genes are generally longer. The length of introns in human genes declines steadily as the rate of expression increases, and the same pattern occurs in *C. elegans*, although it is less pronounced (Fig. 2;

[1]*Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.* [2]*National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. Correspondence should be addressed to F.K. (e-mail: fkondras@ncbi.nlm.nih.gov).*

Table 1). In *C. elegans*, there is also a notable dependence between expression and total exon length, in agreement with previous reports[9–11]; this dependence was much weaker and not statistically significant among human genes (Fig. 2; Table 1).

We also tested the hypothesis that highly expressed genes are under selection for lower intron density (number of introns per unit of coding sequence length) compared with genes expressed at low levels. Taking gene length into account, we determined the average intron density for genes with different expression levels and could detect no clear relationship between intron density and expression (data not shown).

Highly expressed genes from both species were enriched for ribosomal proteins (60% and 40% of the highly expressed genes in human and nematode, respectively). The remaining genes that were highly expressed showed a broad spectrum of functions, although in *C. elegans*, the set of highly expressed genes was enriched for those encoding secreted proteins and structural components of the cuticle, comprising approximately 20% of the highly expressed gene set (see URL listings in Methods). To eliminate the possibility that short introns are primarily characteristic of ribosomal protein genes, we repeated the analysis after removing these genes from the highly expressed gene sets of both species. This did not significantly affect the observed relationship between expression and intron length ($P > 0.5$; Fig. 2).

Pooling EST libraries to estimate the overall expression of human genes could bias estimates of gene expression through the inclusion of normalized and tumor libraries, the overrepresentation of certain tissues or both. To address this possibility, we estimated the expression of human genes from a collection of non-normalized EST libraries obtained from healthy brain tissue. The limited number of ESTs in this collection hampered extensive analysis; nevertheless, highly expressed genes that were identified using the brain-specific library contained significantly shorter introns than genes expressed at lower levels (data not shown).

Taken together, these observations indicate that intron length may be subject to natural selection driven by the advantage incurred by minimizing the cost of transcription. This is compatible with the observation that differences in intron length between genes expressed at high and low levels are greater in humans than in nematodes because humans have, on average, much longer introns. In human genes that are highly expressed, the average intron length is only about three times as large as that in nematode genes that are highly expressed, whereas for genes expressed at low levels the difference between the two species was approximately 20-fold.

These observations raise questions concerning the nature of the selection acting on intron length. Short introns may be an ancestral feature of highly expressed genes, and negative (purifying) selection may act to prevent their growth. Alternatively, intron length in highly expressed genes may have decreased over evolutionary time under positive selection. The likelihood of each interpretation depends on the evolutionary lability of gene expression in general. For example, genes encoding ribosomal proteins or major cytoskeletal proteins, which are prominent among highly expressed genes in both humans and nematodes

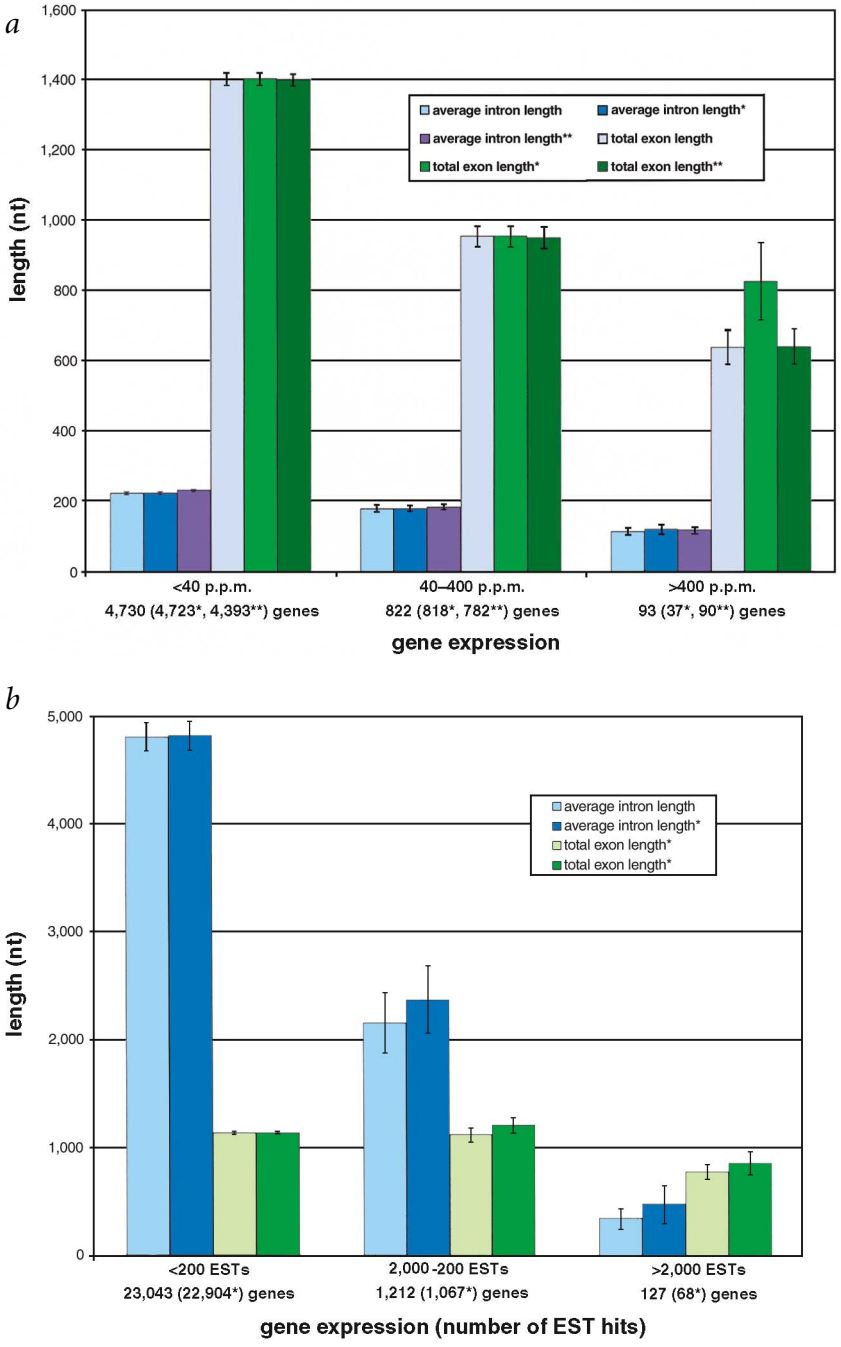

Fig. 2 Comparison of average intron lengths and total exon lengths in genes with high and low expression. ***a***, Genes from *C. elegans*. ***b***, Genes from *H. sapiens*. Error bars indicate the 95% confidence intervals. We subdivided the genes for analysis as follows: *, excluding ribosomal protein genes; **, excluding genes expressed in the germline.

Table 1 • Comparison of average intron length and total exon length in genes expressed at different levels[a]

	Average intron length			Total exon length		
	All genes	Excluding ribosomal protein genes	Excluding genes expressed in the germline	All genes	Excluding ribosomal protein genes	Excluding genes expressed in the germline
***C. elegans* genes**						
<40 p.p.m. versus	224.0 ± 3.5	224.1 ± 3.5	231.1 ± 3.7	1,401.0 ± 18	1,401.1 ± 18	1,397.8 ± 19.1
40–400 p.p.m.	179.7 ± 7.5 $P < <10^{-10}$	179.7 ± 7.6 $P < <10^{-10}$	183.6 ± 7.9 $P < <10^{-10}$	953.7 ± 29.9 $P < <10^{-10}$	953.2 ± 29.9 $P < <10^{-10}$	949.3 ± 30.8 $P < <10^{-10}$
40–400 versus	179.7 ± 7.5	179.7 ± 7.6	183.6 ± 7.8	953.7 ± 29.9	953.2 ± 29.9	949.3 ± 30.8
>400 p.p.m.	113.7 ± 9.7 $P < 0.0079$	119.2 ± 13.4 $P < 0.0426$	116.4 ± 9.9 $P < 0.0135$	636.5 ± 48.5 $P < 2.8 \times 10^{-6}$	824.1 ± 109.9 $P < 0.180$	639.2 ± 50 $P < 7.9 \times 10^{-6}$
***H. sapiens* genes**						
<200 EST hits versus	4,807.0 ± 50.0	4,818.2 ± 86.0	data not available	1,135.3 ± 32.7	1,138.6 ± 55.9	data not available
200–2,000 EST hits	2,153.2 ± 140.9 $P < <10^{-10}$	2,366.0 ± 158.2 $P < <10^{-10}$		1,122.6 ± 31.9 $P < 0.612$	1,211.0 ± 35.3 $P < 0.0023$	
200–2,000 EST hits versus	2,153.2 ± 140.9	2,366.0 ± 158.2	data not available	1,122.6 ± 31.9	1,211.0 ± 35.3	data not available
>2,000 EST hits	342.5 ± 65.2 $P < <10^{-10}$	476.1 ± 65.5 $P < <10^{-10}$		776.8 ± 7.1 $P < 0.058$	859.6 ± 7.1 $P < 0.0566$	

[a]We used the Mann-Whitney *U*-test to determine significance. The upper and lower bounds are s.e.m.

(see URL listings in Methods), are probably highly expressed throughout the Eukaryota and, accordingly, probably had ancestrally short introns[12]. In contrast, genes with more specific functions may have attained a high expression relatively recently and experienced a subsequent reduction in intron length as the result of positive selection.

The prevalence of transposable elements in the human lineage indicates that the shorter length of introns in highly expressed genes may be due to selection against transposable-element insertion. This explanation seems plausible because, if fitness is a function of intron size, negative selection would most efficiently eliminate large insertions, such as those of transposable elements. We determined the content of transposable elements in introns from 382 highly expressed human genes and a random sample of 382 genes expressed at low levels. An average of 16% and 35% of introns, respectively, contained transposable elements. When the sequences that were related to transposable elements were removed from the introns of the two gene sets, however, the average length of the remaining intron sequences in highly expressed genes was still significantly shorter than that in the genes expressed at low levels (912 and 3,208 bp, respectively). Thus, although there is a substantial difference in the content of transposable elements in introns depending on the expression of a gene, selection against transposable-element insertions does not completely explain differences in intron length between genes with different levels of expression. Apparently, natural selection must also act on mutational events unrelated to transposition through negative selection against insertions, through positive selection for deletions in the introns of highly expressed genes or through both mechanisms.

In principle, mutational biases towards deletion[13–15] combined with an increased mutation rate in highly transcribed genes[16,17] could produce the observed relationship between intron length and expression. Such a combination of mutational biases cannot, however, fully explain the above observations because it can account for shorter introns only in genes that are highly expressed in the germline. In *C. elegans*, a relatively small percentage of genes is actively expressed in the germline[18], and these genes account for only 6% (412 of 5,905) of the genes examined in this study. Exclusion of all germline genes from the data does not significantly affect the results (Fig. 2*a*; Table 1). In addition, the functional features of many genes in the highly expressed sets, such as those encoding cuticle components and other extracellular proteins in the nematode, indicate that their active expression may be required in somatic tissues rather than in the germline.

As there is an apparent selection for shorter introns in highly expressed genes and as many of these genes belong to families of paralogs (see URL listings in Methods) that might have evolved, at least in part, by means of retrotransposition, it seems unexpected that highly expressed genes in humans and nematodes are not enriched for intronless genes. In humans, the proportion of intronless genes among genes expressed at high and low levels is similar: 22.0% and 22.4%, respectively; the corresponding values for the nematode are 3.2% and 1.7%. Thus, it seems that short introns, but not the absence or loss of introns, are favored by natural selection in highly expressed genes. This pattern might be explained, in part, by functional constraints on introns at the level of gene regulation. For example, it has been shown that splicing, which only occurs in transcripts that contain introns, is linked to the nucleocytoplasmic transport of mRNAs[19].

Factors other than the direct cost of transcription might also affect intron length. One possibility is that the energetic cost, rate of splicing or both depends on intron length, with short introns being spliced more efficiently. If such a phenomenon exists, it would likewise result in selection for short introns in highly expressed genes. Short introns have a lower rate of recombination[20,21] and a higher guanine and cytosine content[22,23]. It is not, however, clear to what extent, if at all, selection has acted to regulate intron length in connection with these factors[21,22]. It has been suggested that guanine/cytosine-rich introns are shorter because guanine/cytosine-rich genes tend to be highly expressed[23]. We found no correlation, however, between guanine or cytosine content and expression in human genes (data not shown). Additionally, imprinted genes contain short introns[24], but it remains to be determined whether all of these genes are highly expressed.

Natural selection is thought to be important in the evolution of silent sites in coding regions, resulting in an increased codon bias and, consequently, a high translational rate of highly expressed genes[25]. Here we argue that selection also acts to lower the cost of transcription by reducing the length of introns or by maintaining short introns in highly expressed genes. Future studies might reveal other selective forces acting in a similar direction: for example, selection on splicing efficiency.

Methods

Expression data. We obtained expression data for *C. elegans* from available microarray experiments carried out at different stages of development[8]. Expression data for genes designated 'absent' or 'absent at least once' at a particular developmental stage[8] were discarded and the average expression was then taken for each gene in the present analysis. We included 5,632 genes in the analysis, including those that were both significantly modulated and non-modulated through development[8].

Using the number of EST sequences in databases that align unequivocally to a given gene, we estimated the expression of 24,382 human genes. This method gives a reasonably accurate approximation of expression[26–28]. We compared the set of human sequences that encode proteins with the human EST database using the program BLASTN[29]. We accepted EST hits of >400 nt and with >95% identity to a coding region sequence as matches. If they showed >98% identity, we accepted hits of 100–400 nt, and we discarded hits of <100 nt. Genes (mostly *ab initio* predictions) for which no ESTs were detected were not included in the analyses. We obtained all non-normalized EST libraries from normal human brain tissue with >400 ESTs through the UniLib site at NCBI (NIH, Bethesda, Maryland) using 'non-normalized', 'normal' and 'brain' as keywords. We obtained sequences that encode proteins and the data on the number and length of introns from the nematode and human genome[5,30] entries at NCBI. For human genes, we retained only the longest of all overlapping coding sequences (alternative splice forms) for analysis. To estimate the transposon content of human genes, we analyzed intron sequences of genes with >500 ESTs (382 genes total) using the program RepeatMasker, disregarding regions of low complexity. To measure the transposon content of little-expressed genes, we randomly selected 382 genes with 1–50 EST hits. We obtained 95% confidence intervals for average intron size and total exon size in each expression class using non-parametric bootstrapping with 1,000 replicates per class. For the purpose of functional annotation of highly expressed genes, we compared the respective protein sequences with the nonredundant protein sequence database at the NCBI by running three iterations of the PSI-BLAST program[29].

URLs. UniLib, http://www.ncbi.nlm.nih.gov/UniLib/; complete genomes: ftp://ncbi.nlm.nih.gov/genomes/; RepeatMasker, http://ftp.genome.washington.edu/RM/RepeatMasker.html. Gene annotations are available at ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/expression/.

Acknowledgments

We are grateful to A. Kondrashov, I. Rogozin and A. Feldman for reading the manuscript and P. Bouman, J. Cherry, J. Blumensteil and T. Kim for discussion.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 17 April; accepted 24 June 2002.

1. Ucker, D.S. & Yamamoto, K.R. Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. *J. Biol. Chem.* **259**, 7416–7420 (1984).
2. Izban, M.G. & Luse, D.S. Factor-stimulated RNA polymerase-II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *J. Biol. Chem.* **267**, 13647–13655 (1992).
3. Lehninger, A.L., Nelson, D.L. & Cox, M.M. *Principles of Biochemistry* 615–644 (Worth, New York, 1982).
4. Deutsch, M. & Long, M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**, 3219–3228 (1999).
5. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
6. Ogata, H., Fujibuchi, W. & Kanehisa, M. The size differences among mammalian introns are due to the accumulation of small deletions. *FEBS Lett.* **390**, 99–103 (1996).
7. Moriyama, E.N., Petrov, D.A. & Hartl, D.L. Genome size and intron size in *Drosophila*. *Mol. Biol. Evol.* **15**, 770–773 (1998).
8. Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G. & Brown, E.L. Genomic analysis of gene expression in *C. elegans*. *Science* **290**, 809–812 (2000).
9. Eyre-Walker, A. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol. Biol. Evol.* **13**, 864–872 (1996).
10. Duret, L. & Mouchiroud, D. Expression pattern and, suprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **96**, 4482–4487 (1999).
11. Coghlan, A. & Wolfe, H. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16**, 1131–1145 (2000).
12. Nixon, J.E. *et al.* A spliceosomal intron in *Giardia lamblia*. *Proc. Natl Acad. Sci. USA* **99**, 3701–3705 (2002).
13. Ophir, R. & Graur, D. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**, 191–202 (1997).
14. Petrov, D.A., Lozovskaya, E.R. & Hartl, D.L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349 (1998).
15. Robertson, H.M. The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletion and intron gains and losses. *Genome Res.* **10**, 192–203 (2000).
16. Boulikas, T. Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J. Mol. Evol.* **35**, 156–180 (1992).
17. Sullivan, D.T. DNA excision repair and transcription: implications for genome evolution. *Curr. Opin. Genet. Dev.* **5**, 786–791 (1995).
18. Reinke, V. *et al.* A global profile of germline gene expression in *C. elegans*. *Mol. Cell* **6**, 605–616 (2000).
19. Zhou, Z. *et al.* The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* **407**, 401–405 (2000).
20. Carvalho, A.B. & Clark, A.G. Intron size and natural selection. *Nature* **401**, 344 (1999).
21. Comeron, J.M. & Kreitman, M. The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. *Genetics* **156**, 1175–1190 (2000).
22. Hurst, L.D., Brunton, C.F.A. & Smith, N.G.C. Small introns tend to occur in GC-rich regions in some but not all vertebrates. *Trends Genet.* **15**, 437–439 (1999).
23. Carels, N. & Bernardi, G. Two classes of genes in plants. *Genetics* **154**, 1819–1825 (2000).
24. Hurst, L.D. & McVean, G. Imprinted genes have few and short introns. *Nature Genet.* **12**, 234–237 (1996).
25. Akashi, H. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* **11**, 660–666 (2001).
26. Okubo, K. *et al.* Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2**, 173–179 (1992).
27. Lee, N.H. *et al.* Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc. Natl Acad. Sci. USA* **92**, 8303–8307 (1995).
28. Bortoluzzi, S. & Danieli, G.A. Towards an *in silico* analysis of transcription patterns. *Trends Genet.* **15**, 118–119 (1999).
29. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
30. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).