joint project between EMBL–EBI and the Sanger Centre called 'Ensembl' and at Oak Ridge National Laboratory by the Genome Annotation Consortium. The Mouse Genome Sequence project at The Jackson Laboratory is involved in the overall process of data integration.

## Summary

To many, the recommendations of the NIH priority setting meetings seemed impossibly ambitious. However, as the new century unfolds, it is clear that many of these recommendations will soon become realities. Working-draft mouse genomic sequence, combined with transcript-map and full-length cDNA sequence will be a powerful tool in gene identification and annotation, both in the mouse and human genomes. A wealth of new mouse mutations and traits will be identified and mapped and the location of their human equivalents inferred from homology maps. These mutations will comprise a functional map of the genome that can be correlated with the sequence map and expression data to provide a rich resource of biological information. Access to the mutant mice will ensure that a new generation of pathobiologists will be able to characterize better models of human disease. More and more, the mouse could become the measure of man.

## Acknowledgements

M. Strivens and S. Brown at MRC Harwell made helpful comments on the manuscript. We also thank all those workers who helped with unpublished or pre-publication information. This work was supported by the UK Medical Research Council and the US Department of Public Health Grant P01 CA75719.

**References**

1 Battey, J. *et al.* (1999) An action plan for mouse genomics. *Nat. Genet.* 21, 73–75
2 Nusbaum, C. *et al.* (1999) A YAC-based physical map of the mouse genome. *Nat. Genet.* 22, 388–393
3 McCarthy, L.C. *et al.* (1997) A first-generation whole genome-radiation hybrid map spanning the mouse genome. *Genome Res.* 7, 1153–1161
4 Van Etten, W.J. *et al.* (1999) Radiation hybrid map of the mouse genome. *Nat. Genet.* 22, 384–387
5 Lindblad-Toh, K. *et al.* (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* 24, 381–386
6 Rinchik, E.M. and Carpenter, D.A. (1999) N-ethyl-N-nitrosourea mutagenesis of a 6- to 11-cM subregion of the Fah–Hbb interval of mouse chromosome 7: completed testing of 4557 gametes and deletion mapping and complementation analysis of 31 mutations. *Genetics* 152, 373–383
7 Justice, M.J. *et al.* (1997) Using targeted large deletions and high-efficiency N-ethyl-N-nitrosourea mutagenesis for functional analyses of the mammalian genome. *Methods* 13, 423–436
8 Zheng, B. *et al.* (1999) A system for rapid generation of coat color-tagged knockouts and defined chromosomal rearrangements in mice. *Nucleic Acids Res.* 27, 2354–2360
9 Zheng, B. *et al.* (1999) Engineering a mouse balancer chromosome. *Nat. Genet.* 22, 375–378
10 Brown, S.D. and Nolan, P.M. (1998) Mouse mutagenesis – systematic studies of mammalian gene function. *Hum. Mol. Genet.* 7, 1627–1633
11 Justice, M.J. *et al.* (1999) Mouse ENU mutagenesis. *Hum. Mol. Genet.* 8, 1955–1963
12 Wiles, M.V. *et al.* (2000) Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells. *Nat. Genet.* 24, 13–14
13 Matin, A. *et al.* (1999) Susceptibility to testicular germ-cell tumours in a 129.MOLF-Chr 19 chromosome substitution strain. *Nat. Genet.* 23, 237–240
14 Williams, R.W. (1999) A targeted screen to detect recessive mutations that have quantitative effects. *Mamm. Genome* 10, 734–738
15 Nadeau, J.H. *et al.* (2000) Analysing complex genetic traits with chromosome substitution strains. *Nat. Genet.* 24, 221–225
16 Marschall, S. *et al.* (1999) Reliable recovery of inbred mouse lines using cryopreserved spermatozoa. *Mamm. Genome* 10, 773–776
17 Thornton, C.E. *et al.* (1999) Large numbers of mice established by *in vitro* fertilization with cryopreserved spermatozoa: implications and applications for genetic resource banks, mutagenesis screens, and mouse backcrosses. *Mamm. Genome* 10, 987–992
18 Wakayama, T. and Yanagimachi, R. (1998) Development of normal mice from oocytes injected with freeze-dried spermatozoa. *Nat. Biotechnol.* 16, 639–641
19 Liu, X.Y. *et al.* (1999) The gene mutated in bare patches and striated mice encodes a novel 3beta-hydroxysteroid dehydrogenase. *Nat. Genet.* 22, 182–187
20 Mallon, A-M. *et al.* Comparative genome sequence analysis of the Bpa/Str region in mouse and man. *Genome Res.* (in press)
21 Lund, J. *et al.* (1999) Sequence-ready physical map of the mouse chromosome 16 region with conserved synteny to the human velocardiofacial syndrome region on 22q11.2. *Mamm. Genome* 10, 438–443



# tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes

**Laurent Duret**
duret@biomserv.univ-lyon1.fr

Laboratoire 'Biométrie et Biologie Evolutive', UMR 5558. Bâtiment 711, Université Claude Bernard-Lyon 1, 69622 Villeurbanne Cedex, France.

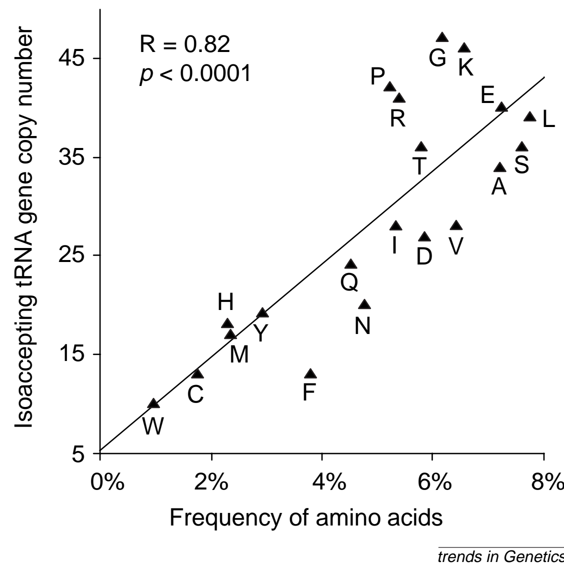
Although they encode the same amino acids, synonymous codons are not all used at the same frequency. Such codon usage biases occur in most species and could be the result of mutational biases, natural selection acting on silent changes in DNA, or both[1]. Selection on synonymous codon positions is thought to lead to a co-adaptation of codon usage and tRNA content to optimize the efficiency of translation. Such a selective pressure to reduce the cost of translation is expected to be stronger for genes that are expressed at high levels. In agreement with that model, in some unicellular organisms such as *Escherichia coli* or *Saccharomyces cerevisiae*, codons that are used preferentially correspond to the most abundant tRNA species[2,3], and there is a positive correlation between codon usage bias and the level of gene expression[1].

In multicellular eukaryotes, there are very few experimental data on tRNA abundance. However, the limited data available in *Drosophila*[4] showed a relationship between codon usage and tRNA abundance[5]. Moreover, in *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*, the frequency of favored codons (i.e. codons that are used preferentially in genes that have a strong codon-usage bias) increases with increasing expression level[6].

With the completion of the sequence of *C. elegans* genome[7], we have the opportunity to study the first complete tRNA collection from a complex, multicellular eukaryote and to assess the relationship with codon usage. tRNA genes in the *C. elegans* genome were searched using the program tRNAscan-SE with the default (eukaryote-specific) parameters[8]

0168-9525/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)02041-2

**FIGURE 1. Amino acid frequency and isoaccepting tRNA genes in *C. elegans***




Correlation between the frequency of amino-acids in the *Caenorhabditis elegans* proteome (weighted according to gene expression level) and the number of isoaccepting tRNA genes.

(T.M. Lowe, pers. commun.). tRNAscan-SE predicted 592 tRNAs and 194 pseudogenes. After visual inspection, 13 borderline tRNAs (scoring between 20 and 30 bits) were judged to be probable pseudogenes. Thus, the revised tally is 579 tRNAs and 207 tRNA-like pseudogenes. tRNAscan-SE was also used in the analyses presented in the *C. elegans* genome publication[7], which gives a different overall tRNA count. We believe that data was incorrectly reported owing to simple error or changes in the final version of the sequence (T.M. Lowe, pers. commun.).

There is a high level of tRNA gene redundancy in the *C. elegans* genome: besides tRNA$^{SeC}$, which is encoded by a single gene, gene copy number varies from ten copies for tRNA$^{Trp}$ to 46 copies for tRNA$^{Gly}$. Although the level of transcription might vary among tRNA genes, one might expect that the cellular tRNA abundance should be correlated with the number of tRNA genes. Indeed, in some unicellular organisms, it has been shown that tRNA gene copy number is an important factor in the determination of intracellular tRNA levels[2,3,9]. To study whether there is also such a gene-dosage effect in multicellular eukaryotes, we analysed the relationship between the number of isoaccepting tRNA genes and the frequency of the 20 amino-acids in the proteome of *C. elegans*. The quantity of amino acids incorporated into newly synthesized proteins depends on expression levels, which vary widely according to the genes. To take this variation into account, the mRNA abundance of each *C. elegans* gene

**FIGURE 2. Favored codons and isoaccepting tRNAs**

| tRNA copy | gene number | RGF | Anticodon | Codon | | RSCU |
|---|---|---|---|---|---|---|
| Gly | **31** | **2.64** | **TCC** | GGA | • | 2.77 |
| | 13 | 1.11 | GCC | GGC | | 0.4 |
| | 3 | 0.26 | CCC | GGG | | 0.15 |
| | 0 | 0.00 | ACC | GGT | | 0.69 |
| Val | 5 | 0.71 | TAC | GTA | | 0.25 |
| | 0 | 0.00 | GAC | GTC | • | 1.56 |
| | 5 | 0.71 | CAC | GTG | | 0.71 |
| | **18** | **2.57** | **AAC** | GTT | | 1.47 |
| Lys | 15 | 0.65 | TTT | AAA | | 0.59 |
| | **31** | **1.35** | **CTT** | AAG | • | 1.41 |
| Asn | **20** | **2.00** | **GTT** | AAC | • | 1.22 |
| | 0 | 0.00 | ATT | AAT | | 0.78 |
| Gln | **18** | **1.50** | **TTG** | CAA | | 1.19 |
| | 6 | 0.50 | CTG | CAG | • | 0.81 |
| His | **18** | **2.00** | **GTG** | CAC | • | 1.23 |
| | 0 | 0.00 | ATG | CAT | | 0.77 |
| Glu | 17 | 0.85 | TTC | GAA | | 0.92 |
| | **23** | **1.15** | **CTC** | GAG | • | 1.08 |
| Asp | **27** | **2.00** | **GTC** | GAC | • | 0.92 |
| | 0 | 0.00 | ATC | GAT | | 1.08 |
| Tyr | **19** | **2.00** | **GTA** | TAC | • | 1.35 |
| | 0 | 0.00 | ATA | TAT | | 0.65 |
| Cys | **13** | **2.00** | **GCA** | TGC | • | 1.28 |
| | 0 | 0.00 | ACA | TGT | | 0.72 |
| Phe | **13** | **2.00** | **GAA** | TTC | • | 1.5 |
| | 0 | 0.00 | AAA | TTT | | 0.5 |
| Ile | 7 | 0.75 | TAT | ATA | | 0.1 |
| | 0 | 0.00 | GAT | ATC | • | 1.75 |
| | **21** | **2.25** | **AAT** | ATT | | 1.16 |
| Met | 17 | | CAT | ATG | | |
| Trp | 10 | | CCA | TGG | | |
| SeC | 1 | | TCA | TGA | | |
| Arg | 7 | 1.02 | TCT | AGA | | 1.56 |
| | 4 | 0.59 | CCT | AGG | | 0.15 |
| | 10 | 1.46 | TCG | CGA | | 0.62 |
| | 0 | 0.00 | GCG | CGC | • | 1.34 |
| | 1 | 0.15 | CCG | CGG | | 0.24 |
| | **19** | **2.78** | **ACG** | CGT | • | 2.09 |
| Leu | 3 | 0.46 | TAG | CTA | | 0.19 |
| | 0 | 0.00 | GAG | CTC | • | 1.92 |
| | 6 | 0.92 | CAG | CTG | | 0.69 |
| | **19** | **2.92** | **AAG** | CTT | • | 1.76 |
| | 4 | 0.62 | TAA | TTA | | 0.19 |
| | 7 | 1.08 | CAA | TTG | | 1.25 |
| Ser | 8 | 1.33 | GCT | AGC | | 0.78 |
| | 0 | 0.00 | AGT | AGT | | 0.55 |
| | 7 | 1.17 | TGA | TCA | | 0.99 |
| | 0 | 0.00 | GGA | TCC | • | 1.39 |
| | 6 | 1.00 | CGA | TCG | | 0.98 |
| | **15** | **2.50** | **AGA** | TCT | | 1.31 |
| Thr | 12 | 1.33 | TGT | ACA | | 0.79 |
| | 0 | 0.00 | GGT | ACC | • | 1.5 |
| | 7 | 0.78 | CGT | ACG | | 0.42 |
| | **17** | **1.89** | **AGT** | ACT | | 1.29 |
| Pro | **32** | **3.05** | **TGG** | CCA | • | 2.82 |
| | 0 | 0.00 | GGG | CCC | | 0.19 |
| | 4 | 0.38 | CGG | CCG | | 0.61 |
| | 6 | 0.57 | AGG | CCT | | 0.38 |
| Ala | 8 | 0.94 | TGC | GCA | | 0.66 |
| | 0 | 0.00 | GGC | GCC | • | 1.4 |
| | 4 | 0.47 | CGC | GCG | | 0.31 |
| | **22** | **2.59** | **AGC** | GCT | • | 1.63 |

trends in Genetics

Relationships between favored codons and isoaccepting tRNAs having the highest gene copy number. Codons predicted to be decoded by the major tRNA gene (in bold) are indicated by an arrow (dashed arrows denote the preferential pairing of I with U or C over A). The relative gene frequency of tRNA genes (RGF) is the observed frequency of an isoacceptor tRNA gene in *C. elegans* genome divided by the frequency expected if all isoacceptor tRNA genes for that amino acid were equally frequent in the genome. The relative synonymous codon usage (RSCU) is the observed frequency of a codon divided by the frequency expected if all synonyms for that amino acid were used equally. RSCU was measured in short (less than 333 codons) highly expressed genes (data from Ref. 6). Favored codons[13] are indicated by a dot.

(15 425 predicted complete coding regions) was estimated by counting the number of matching Expressed Sequence Tags (ESTs) available in GenBank, as described previously[6]. This method gives only a rough estimate of gene expression level, because these ESTs have been sequenced from cDNA libraries (prepared from embryo and adult whole organisms) that have been partly normalized (Y. Kohara, unpublished). Despite this, we found a strong correlation between the number of tRNA genes and the frequency of amino-acids among *C. elegans* proteins, weighted according to their expression level (Fig. 1). As expected, this correlation is significantly stronger for highly expressed genes ($R=0.82$, $p<0.0001$, $N=1631$) compared with those that do not match any EST (i.e. genes with very low expression level; $R=0.61$, $p=0.0042$, $N=9409$). If the cellular tRNA abundance were not related to the number of tRNA gene, we would not expect any correlation between the frequency of amino acids and the number of tRNA genes. Therefore, these observations strongly support that in *C. elegans*, as in unicellular organisms, intracellular tRNA levels are mainly determined by gene copy number.

Interestingly, it has been noted that the tRNA gene density is higher on the X chromosome, compared with autosomes[7]. Indeed, there are 15.7 tRNA genes per Mb on the X chromosome, but only 3.4–4.4 tRNA genes per Mb on the autosomes – what is the reason for this? Evolutionary theory predicts that sex chromosomes evolved from an autosomal pair, and once X–Y recombination ceased, Y-linked genes were progressively inactivated and obliterated (ultimately up to the final loss of the Y-chromosome, as in *C. elegans*). As an adaptive response to the loss of Y-linked genes, homologous genes on the X-chromosome were up-regulated and subsequently subject to dosage compensation[10]. We propose that, in *C. elegans*, the up-expression of X-linked tRNA genes was achieved by gene duplication, which has led to the present excess of tRNA genes on the X chromosome compared with autosomes.

Gene copy numbers vary greatly among isoaccepting tRNAs. For example, among the 42 $tRNA^{Pro}$ genes, 32 (76%) contain the TGG anticodon (Fig. 2). Is this variation related to a bias in codon usage? To answer this question, we computed the relative gene frequency (RGF) of each isoacceptor tRNA in the genome and the relative synonymous codon usage (RSCU) in highly expressed genes. There is a highly significant correlation between RGF and the RSCU of complementary codons ($R=0.54$, $p<0.0001$). It should be noticed that this correlation reflects only partially the real co-adaptation of tRNA abundance and codon usage because the same tRNA can decode several codons. Although we have no experimental data on base modifications in *C. elegans* tRNAs (except for tRNA Leu-AAG, with an inosine modification at the first anticodon position[11]), it is possible to predict the codons decoded by the different anticodons according to the classical rules[12]: (1) G–U wobble pairing; (2) an I at the first anticodon position produces a preference for U or C over A; and (3) As at the first anticodon position are predicted as Is, because an unmodified A has been found only in a few exceptional cases. This analysis revealed that all the favored codons are decoded by the isoaccepting tRNA that has the highest gene copy number (Fig. 2). In the three cases where there are two favored codons (Arg, Leu and Ala), both codons are decoded by the major isoaccepting tRNA. Moreover, in all cases where an I is predicted at the first anticodon position, favored codons end in C and/or U. Finally, for eight of the nine duets, the favored codon is the direct complement of the major isoaccepting tRNA. In conclusion, codon-usage biases and tRNA gene redundancy in *C. elegans* genome clearly reflect a co-adaptation of tRNA content and codon usage for the optimal translation of the pool of highly expressed genes.


## Acknowledgements

I warmly thank T.M. Lowe for providing me the results of his analyses of tRNA genes contents in *C. elegans* genome. I thank M. Gouy for his helpful comments and G. Marais for his suggestion about the relationship between the frequency of tRNA genes and dosage compensation on the X chromosome. This work was supported by the Centre National de la Recherche Scientifique.



**References**

1 Sharp, P.M. *et al.* (1993) Codon usage: mutational bias, translational selection, or both? *Biochem. Soc. Trans.* 21, 835–841
2 Dong, H. *et al.* (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* 260, 649–663
3 Kanaya, S. *et al.* (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238, 143–155
4 White, B.N. *et al.* (1973) Analysis of tRNAs during the development of *Drosophila. Dev. Biol.* 33, 185–195
5 Moriyama, E.N. and Powell, J.R. (1997) Codon usage bias and tRNA abundance in *Drosophila. J. Mol. Evol.* 45, 514–523
6 Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis. Proc. Natl. Acad. Sci. U. S. A.* 96, 4482–4487
7 The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018
8 Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964
9 Percudani, R. *et al.* (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae. J. Mol. Biol.* 268, 322–330
10 Jegalian, K. and Page, D.C. (1998) A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* 394, 776–780
11 Tranquilla, T.A. *et al.* (1982) Sequences of four tRNA genes from *Caenorhabditis elegans* and the expression of *C. elegans* tRNALeu (anticodon IAG) in *Xenopus* oocytes. *Nucleic Acids Res.* 10, 7919–7934
12 Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2, 13–34
13 Stenico, M. *et al.* (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.* 22, 2437–2446


## Author's correction

Deng, X-W. *et al.* (2000) Unified nomenclature for the COP9 signalosome and its subunits: an essential regulator of development. *Trends Genet.* 16, 202–203

An error occured during the publication of the letter by Deng *et al.* in the May issue. The correct citation for this letter should be:

Deng, X-W., Dubiel, W., Wei, N., Hofmann, K., Mundt, K., Colicelli, J., Kato, Jy., Naumann, M., Segal, D., Seeger, M., Carr, A., Glickman, M., Chamovitz, D.A.

PII: S0168-9525(00)02071-0