

Top

The equations in this article are [MathML-enabled](#). Install [MathPlayer](#) for free to take advantage of this.

Abstract

Research

Open Access

Background

Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species

Results

Discussion

Conclusion

Makedonka Mitreva¹ ✉, Michael C Wendl¹ ✉, John Martin¹ ✉, Todd Wylie¹ ✉, Yong Yin¹ ✉, Allan Larson² ✉, John Parkinson³ ✉, Robert H Waterston⁴ ✉ and James P McCarter^{1,5} ✉

¹ Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA

² Department of Biology, Washington University, St. Louis, Missouri 63130, USA

³ Hospital for Sick Children, Toronto, and Departments of Biochemistry/Medical Genetics and Microbiology, University of Toronto, M5G 1X8, Canada

⁴ Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

⁵ Divergence Inc., St Louis, Missouri 63141, USA

Materials and methods

Additional data files

Acknowledgements

✉ author email ✉ corresponding author email

References

Genome Biology 2006, **7**:R75 doi:10.1186/gb-2006-7-8-r75

Subject areas: Molecular biology, Evolution, Genetics

The electronic version of this article is the complete one and can be found online at:

<http://genomebiology.com/2006/7/8/R75>

Received: 20 April 2006
Revisions received: 30 June 2006
Accepted: 14 August 2006
Published: 14 August 2006

© 2006 Mitreva et al.; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background

Codon usage has direct utility in molecular characterization of species and is also a marker for molecular evolution. To understand codon usage within the diverse phylum Nematoda, we analyzed a total of 265,494 expressed sequence tags (ESTs) from 30 nematode species. The full genomes of *Caenorhabditis elegans* and *C. briggsae* were also examined. A total of 25,871,325 codons were analyzed and a comprehensive codon usage table for all species was generated. This is the first codon usage table available for 24 of these organisms.

Results

Codon usage similarity in Nematoda usually persists over the breadth of a genus but then rapidly diminishes even within each clade. *Globodera*, *Meloidogyne*, *Pristionchus*, and *Strongyloides* have the most highly derived patterns of codon usage. The major factor affecting differences in codon usage between species is the coding sequence GC content, which varies in nematodes from 32% to 51%. Coding GC content (measured as GC3) also explains much of the observed variation in the effective number of codons ($R = 0.70$), which is a measure of codon bias, and it even accounts for differences in amino acid frequency. Codon usage is also affected by neighboring nucleotides (N1 context). Coding GC content correlates strongly with estimated noncoding genomic GC content ($R = 0.92$). On examining abundant clusters in five species, candidate optimal codons were identified that may be preferred in highly expressed transcripts.

Conclusion

Evolutionary models indicate that total genomic GC content, probably the product of directional mutation pressure, drives codon usage rather than the converse, a conclusion that is supported by examination of nematode genomes.

Background

Utilization of the degenerate triplet code for amino acid (AA) translation is neither uniform nor random. In particular, there are distinct

Genome Biology
Volume 7
Issue 8

Viewing options:

- Abstract
- Full text
- PDF (478KB)
- Additional files

Associated material:

- PubMed record

Related literature:

- Articles citing this article on BioMed Central on Google Scholar on PubMed Central
- Other articles by authors
 - on Google Scholar
 - Mitreva M
 - Wendl MC
 - Martin J
 - Wylie T
 - Yin Y
 - Larson A
 - Parkinson J
 - Waterston RH
 - McCarter JP
 - on PubMed
 - Mitreva M
 - Wendl MC
 - Martin J
 - Wylie T
 - Yin Y
 - Larson A
 - Parkinson J
 - Waterston RH
 - McCarter JP
- Related articles/pages on Google on Google Scholar on PubMed
- Evaluation of this article in F1000 Biology

Tools:

- Download citation(s)
- Download XML
- Email to a friend
- Order reprints
- Post a comment
- Sign up for article alerts

Post to:

- Citeulike
- Connotea
- Del.icio.us
- Digg
- Facebook

patterns among different species and genes. Such patterns can readily be characterized by codon usage, namely the observed percentage occurrence with which each codon is used to encode a given AA. This measure has direct utility in molecular characterization of a species in that it enables efficient degenerate and nondegenerate primer design for cross-species gene cloning, open reading frame determination, and optimal protein expression [1]. Such tools are particularly important with respect to species for which limited molecular information exists. Codon usage also serves as an indicator of molecular evolution [2]. Codon usage bias, namely the degree to which usage departs from uniform use of all available codons for an AA, can be influenced by a number of evolutionary processes. The guanine and cytosine (GC) versus adenine and thymine (AT) composition of the species' genome, probably the product of directional mutation pressure [3,4], is a key driver of both codon usage and AA composition [5,6]. Other factors that influence codon usage may include the relative abundance of isoaccepting tRNAs [7-9], especially for highly expressed mRNAs that require translational efficiency [10,11], presence of mRNA secondary structure [12,13], and facilitation of correct co-translational protein folding [14]. Codon usage appears not to be optimized to minimize the impact of errors in translation and replication [15].

Nematodes are a highly abundant and diverse group of organisms that exploit niches from free-living microbivory to plant and animal parasitism. Molecular phylogenies divide nematodes into five major named and numbered clades within which parasitism has arisen multiple times [16]: Dorylaimia (clade I), Enoplia (clade II), Spirurina (clade III), Tylenchina (clade IV), and Rhabditina (clade V). Following the sequencing of the complete genome of the model nematode *Caenorhabditis elegans* [17], we have begun to catalog the molecular diversity of nematode genomes through the generation of over 250,000 expressed sequence tags (ESTs) from more than 30 nematode species (including 28 parasites) in four clades. Gene expression analyses for several medically and economically important parasites such as filarial, hookworm, and root knot nematode species have been completed [18-23] (for reviews [24,25]). Moreover, we recently conducted a meta-analysis of partial genomes across the whole phylum with a focus on the conservation and diversification of encoded protein families [26]. Project information is maintained on several online resources [27-30].

Now, in the most extensive such study yet performed for any phylum, we extend the above analyses with a comprehensive survey of observed codon usage and bias based on nearly 26 million codons in 32 species of the Nematoda. Because of its completed genome, *C. elegans* has been the primary species utilized in nematode codon usage studies [31-34]. Our findings provide more complete information for *Caenorhabditis* based on all 41,782 currently predicted proteins in *C. elegans* and *C. briggsae* [35]. Studies for other nematode species have been more limited. Codon usage has been tabulated for a number of parasitic nematodes including filarial species *Brugia malayi*, *Onchocerca volvulus*, *Wuchereria bancrofti*, *Acanthocheilonema viteae*, *Dirofilaria immitis* [36-39], *Strongyloides stercoralis* [40], *Ascaris suum* [41], *Ancylostoma caninum*, and *Necator americanus* [42]. Although Fadiel and coworkers [39] used up to 60 genes per species, sample sizes in the other studies were quite small, typically fewer than 10 representative genes and 5,000 codons per species. In the present study we used an average of 2,350 genes and 270,000 codons per species for the 30 non-*Caenorhabditis* species. Our results provide the first codon usage tables for 24 of these organisms. Web available automated codon usage databases compiled from GenBank [43] lack almost all of this information because they rely only on full-length protein coding gene sequence submissions rather than the EST data used here.

In analyzing codon distribution in Nematoda, we describe how average usage varies between species and across the phylum. For instance, it has been shown that there is a level of conservation in codon distribution between 'closely' related nematodes such as *Brugia malayi* and *B. pahangi* [37] and *Brugia* and *Onchocerca* [38]. These relationships do not appear to extend over greater evolutionary distances, for instance between *Onchocerca* and *Caenorhabditis* [36]. The evolutionary distance at which conservation of codon usage diminishes has not previously been established [32]. Here we show that codon usage similarity in Nematoda is a relatively short-range phenomenon, generally persisting over the breadth of a genus but then rapidly diminishing within each clade. We also show that the major factor affecting differences in mean codon usage between distantly related species is the coding sequence GC as compared with AT content. GC content also explains much of the observed variation in the effective number of codons, a measure of codon bias, and even differences in AA frequency.

Results

Determination of codon usage patterns and amino acid composition

Extensive nucleotide sequence data are now available for many nematode species, largely because of recent progress using genomic approaches [25,44]. To obtain a better understanding of codon usage and AA composition within the phylum Nematoda, we analyzed a total of 265,494 EST sequences originating from 30 nematode species. The ESTs define 93,645 clusters or putative genes, with 208-9,511 clusters per species (Table 1) [26]. Table 1 also provides two letter codes for the nematode species used throughout the remainder of the report. We used prot4EST, a translation prediction pipeline optimized for EST datasets [45], to generate protein predictions. To reduce noise derived from poor translations, our analysis considered only the longest open reading frame (ORF) translations with strong supporting evidence in the form of similarity to known or predicted proteins (BLASTX cutoff $1 \times e^{-8}$) and retained only the polypeptide aligned portion of the nucleotide sequence. About 75% of the clusters met these criteria, yielding 8,080,057 codons originating from species other than *Caenorhabditis*, and 25,871,325 total codons from all 32 species including available predictions from *C. elegans* and *C. briggsae*. The 18 AA residues with redundant codons gave a total of $(18) \times C_{32,2} = 496$ comparisons of codon usage between species. Comprehensive tables of AA composition (Tables 2 and 3) and codon usage (Table 4) for all 32 Nematoda species studied are provided. Below we use these tables to examine, first, variation in AA composition and its relationship to GC content and, second, codon usage and its relationship to GC content.

Table 1. Summary of sequences used by nematode species

Table 2. Amino acid composition (%) of translations by nematode species

Table 3. Amino acid composition (%) of translations from Nematoda and four reference species

Table 4. Codon usage of translations by nematode species

To examine these variables independent of species relatedness, correlations were calculated using phylogenetically independent contrasts (see Materials and methods, below). The variances of the contrasts were computed for each character as a measure of the variance accumulating per unit branch length. The branch lengths were estimated from the maximum likelihood phylogeny assuming a molecular clock (Figure 1); by this criterion, the tips of the tree are all equidistant in branch length from its root. Computed contrasts were

plotted in all figures representing pair-wise comparisons, and the correlation coefficients were calculated from the paired contrasts. This method is robust to changes in molecular clock assumptions. (Trees calculated without the assumption of a molecular clock are similar in topology but differ in rooting, and branch lengths vary according to amount of base substitution in the 18S rRNA; the clock-based tree provides branch lengths that should estimate most closely the relative durations of branches in evolutionary time. Because independent contrasts are influenced mainly by relative branch lengths, our results should be robust to alternative placements of the root.)

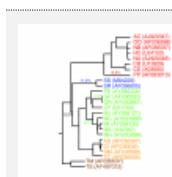


Figure 1. Maximum likelihood (ML) analysis of 18S ribosomal RNA from 25 nematode species. The ML calculation assumes a molecular clock; thus, the tips of the tree are all equidistant, in branch length, from its root. This model of base substitution allows the expected frequencies of the four bases to be unequal, and different rates of evolution at different sites are allowed. The numbers indicate reconstruction of percentage changes in overall codon usage on this phylogenetic topology (see Codon usage patterns and relationships to sampling method, nematode phylogeny, and GC content [under Results]). A distance matrix of D values corrected for non-additivity $[1 - \text{antilog}(-D)] \times 100$ was partitioned on the topology using the cyclic neighbor-joining algorithm, as illustrated by Avise [82]. Approximate percentage change in overall codon usage is indicated for five branches inferred to have undergone 5% or more divergence from an ancestral nematode pattern. This analysis identified genera *Globodera*, *Meloidogyne*, *Pristionchus*, and *Strongyloides* as having the most highly derived patterns of codon usage, and the remaining species as having relatively little net divergence from an ancestral nematode pattern. Definitions of species two letter codes are provided in Table 1; GenBank accession numbers are listed on right. Clades V are shown in red, IVa in blue, IVb in green, III in yellow, and I in brown.

Amino acid composition of nematode proteins and relationship to GC content

AA composition of predicted proteins in nematodes varies among species within a narrow window and is similar to that observed in other organisms (Tables 2 and 3). (Standard deviations in AA usage among nematodes range from 5% to 15% of mean usage, and mean nematode AA usage differs from the mean of four representative organisms by an average of 8%.) Across nematodes, Leu is the most common AA (8.8% of all codons) and Trp the least common (1.1%). Eight AAs contribute an average of more than 6% each to AA content (Ile, Gly, Val, Glu, Ala, Lys, Ser, and Leu); these AAs are also among the most common in the proteomes of other representative species, including humans (Table 3). As in other taxa [46], nematodes show a correlation between AA usage and the degree of codon degeneracy ($R = 0.72$).

In nematodes, coding sequence GC content, derived from our EST clusters, varies from 32% to 51% (Table 1) among species, with a mean of $43.6 \pm 5.9\%$. The distribution is biphasic, with a peak at 36% GC and a second peak at 48%. *Strongyloides* (SS and SR), *Meloidogyne* (MI, MJ, and so on), and filarial parasites (BM, DI, and OV) are the most AT rich (low GC); and NB, PP, and cyst nematodes (GP, GR, and HG) are the most GC rich (approximately 50%). The variation observed in AA composition among species shows a clear relationship to the species' coding sequence GC content. The frequency of AAs encoded by WVN codons (AA, AT, TA, or TT in the first and second nucleotide positions; Asn, Ile, Lys, Try, Phe, and Met) decreases with increasing coding sequence GC content (Figure 2a), whereas the proportion of AAs encoded by SSN codons (GG, GC, CG, and CC; Ala, Arg, Pro, and Gly) increases with higher coding sequence GC content (Figure 2b), and these relationships remain even after removing the effect of evolutionary relationships using phylogenetically independent contrasts. Among AAs, the most uniform and precipitous decrease with increasing GC content was seen with Ile and Tyr whereas the most uniform and rapid increase with higher GC content was seen with Ala and Arg. The trend is less pronounced for other AAs (flatter slope, lower R value). Thr, encoded by four GC/AT 'balanced' codons (ACN), exhibits no change in its frequency with changing GC content (data not shown).

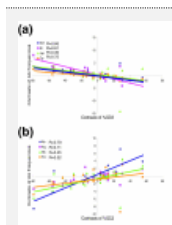


Figure 2. Correlation between phylogenetically independent contrasts of coding sequence GC3 content and AA usage for 25 nematode species. (a) AAs lysine (Lys), isoleucine (Ile), asparagine (Asn), and tyrosine (Tyr) are used less frequently as the species' coding sequence GC3 content increases. (b) AAs alanine (Ala), glycine (Gly), arginine (Arg), and proline (Pro) are used more frequently as the coding sequence GC3 content increases. AA, amino acid.

Base composition by codon position in nematode transcripts and relationship to GC content

Codon usage in nematode species was examined by several methods, including comparison of base usage by position (1-3) over all AAs and comparison of codon usage within each AA. Over all AAs, purine (AG) and pyrimidine (TC) usage in positions 1, 2, and 3 is remarkably uniform between species, favoring purines in position 1 (AG $59.6 \pm 1.5\%$), near equal usage in position 2 (AG $50.0 \pm 0.8\%$), and pyrimidines in position 3 (AG $47.9 \pm 1.5\%$; Additional data file 1). Similar values were observed in *Schistosoma mansoni* (AG 61%, 53%, and 48% in positions 1, 2, and 3, respectively) [1]. GC versus AT usage also varies by position but with much greater variance, with near equal usage in position 1 (50.3% GC) and lower GC usage in positions 2 and 3 (39.1 and 41.4%, respectively), mainly due to greater use of G in position 1 and T in positions 2 and 3 [4].

Additional file 1.

Format: XLS Size: 23KB [Download file](#)

This file can be viewed with: [Microsoft Excel Viewer](#)

The variation observed in GC usage by codon position among species exhibits a clear relationship to the species' overall coding sequence GC content. Not surprisingly, both GC1 and GC2 composition increase with higher coding sequence GC3 content (Figure 3). Specifically, species with high AT content like root-knot *Meloidogyne* species (MI, MJ, and so on) and filarial worms (BM, DI, and OV) [38,39] are biased toward codons terminating in A or T, whereas species with higher GC content such as NB, PP, cyst nematodes, and whipworms (TM and TV) prefer codons ending with G or C. Differences in calculated GC composition by codon position (1-3) between species are determined both by the species' AA usage (as described above) and the codons used for each AA. For example, Cys was encoded by TGI as much as 85% of the time for the AT-rich *Strongyloides* genomes, whereas TGC was used up to 60% of the time in GC-rich genomes such as NB, PP, and HG. To compare codon usage more systematically for individual AAs between species, we employed a statistical approach (described

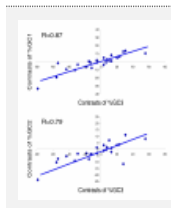


Figure 3. Correlation between phylogenetically independent contrasts of the third position GC content (GC3) and that of the first (GC1) and second (GC2) codon positions for 25 nematode species.

Codon usage patterns and relationships to sampling method, nematode phylogeny, and GC content

Similarity in codon usage was quantified and reported as D_{100} values for each species and AA compared [47,48] (matrix of D_{100} values for each species and AA compared is available in Additional data file 2).

Additional file 2.

Format: XLS Size: 46KB [Download file](#)

This file can be viewed with: [Microsoft Excel Viewer](#)

Because analyses of all but two of the nematode species were based on EST-derived partial genomes [26], comparisons were performed to estimate the differences in codon usage pattern that could be expected using EST collections versus gene predictions derived from a fully assembled and annotated genome. Using *C. elegans*, parallel analyses were performed using either all 22,254 predicted gene products or two EST datasets (*CE-A* and *CE-B*) each comprising 10,000 ESTs. Clustering and peptide predictions were performed using the same algorithms as for the other 30 species. The average D_{100} value for the comparison of codon usage pattern between the *CE-A* and *CE-B* datasets was 0.18, which was not statistically different at the $P < 0.05$ threshold and less than the D_{100} value of the *C. elegans* to *C. briggsae* comparison (0.40). Comparing the *CE-A* and *CE-B* datasets to the genome-derived full gene set for *C. elegans* yielded average D_{100} values of 0.67 and 0.26, respectively. At a practical level, the calculated use of the average codon in *C. elegans* based on *CE-A* and *CE-B* differs from that based on prediction from the whole genome by just $3.4 \pm 2.3\%$ and $2.0 \pm 1.5\%$, respectively. Therefore, although differences in calculated codon usage using partial versus whole genome data are modest enough to make EST-derived codon usage data highly informative, care must be taken not to over-interpret minor differences in D_{100} values because such differences are probably within the range of sampling error (see Discussion, below). However, such uncertainty around small differences in D_{100} values does not alter the major trends that we describe.

The 16 intragenus comparisons of species sharing the same genus name (*Ancylostoma*, *Caenorhabditis*, *Strongyloides*, *Globodera*, *Meloidogyne*, *Ascaris*, and *Trichuris*) all have low D_{100} values, with a mean of 0.14 ± 0.11 (median 0.09, range 0.02-0.40), indicating very similar patterns of codon usage among species within the same genera. By contrast, the 480 comparisons beyond named genera vary greatly, with a mean D_{100} value of 8.10 ± 7.46 (median 5.26, range 0.08-40.56). Low D_{100} values do sometimes extend to comparisons among genera. For instance, relatively low D_{100} values (0.08-1.94) are observed within the following: order Haemonchidae (*HC*, *OO*, and *TD*); subfamily Heteroderinae (*GP*, *GR*, and *HG*); superfamily Ascaridoidea (*AS*, *AL*, and *TC*); and superfamily Filarioidea (*BM*, *DI*, and *OV*). However, low D_{100} values are not maintained across family Ancylostomatidae (*NA*, *AC*, and *AY*), family Strongyloididae (*SS*, *SR*, and *PT*), superfamily Tylenchoidea (*PE-MC*), and order Trichocephalida (*TS*, *TM*, and *TV*). Similarity in codon usage, as indicated by low D_{100} values, does not extend to the level of the major clades (I, III, IVb, IVa, and V).

Furthermore, species with very similar GC content, although distantly related, can exhibit extremely similar codon usage (for instance *Ancylostoma caninum* versus *Toxocara canis*, GC = 48%, $D_{100} = 0.79$). Species with the lowest average D_{100} values in one-versus-all comparisons are those closest to the median species GC content, such as *PE* (GC = 46%). Taxa with the highest AT content, such as *Strongyloides* and *Meloidogyne* species, have among the most extreme differences in codon usage when compared with species beyond their genus (median D_{100} values are 15.3 and 9.4, respectively).

Phylogenetic analysis of changes in codon usage using $(1 - \text{antilog}[-D]) \times 100$, interpretable as percentage divergence in overall codon usage (Figure 1), identifies five branches that have accumulated more than 5% change in codon usage. These branches are as follows: the most recent common ancestor of clades III, IVa, and IVb (5.2%); the most recent common ancestor of clade IVa (11.2%); the most recent common ancestor of genus *Meloidogyne* (6.7%); the most recent common ancestor of genus *Globodera* (7.3%); and the lineage represented by *PP* (8.3%). Genera *Globodera*, *Meloidogyne*, *Pristionchus*, and *Strongyloides* therefore represent the most highly derived patterns of codon usage in nematodes, with the remaining species exhibiting less relatively divergence from an ancestral nematode pattern.

Codon bias in nematode transcripts and relationship to GC content

We used the effective number of codons (ENC) to measure the degree of codon bias for a gene [49]. ENC is a general measure of non-uniformity of codon usage and ranges from 20 if only one codon is used for each AA to 61 if all synonymous codons are used equally. The mean ENC across all sampled nematode species is 46.7 ± 5.1 , and many nematodes have ENC values similar to those obtained for various bacteria, yeast, and *Drosophila* species (ENCs of 45-48) [50]. Outliers with low ENC values include *SS* and *SR*, for which transcripts on average utilize only about 35 of 61 available codons. The variation observed in ENC values among species exhibits a clear relationship to the species' overall coding sequence GC3 content ($R = 0.70$ following phylogenetic correction; Figure 4). The correlation confirms that species with lower GC3 content in coding sequence have greater codon usage bias than those with higher GC3. ENC values for nematodes peak at 47-49% GC (data not shown). In addition to comparing species' mean ENC values, we also examined the distribution of ENC values across all transcripts within each species. Although all species have examples of transcripts across nearly the full range of possible ENC values, in species with low GC3 content, such as *SR*, the distribution is shifted toward a lower ENC peak (Additional data file 3).

Additional file 3.

Format: PPT Size: 51KB [Download file](#)

This file can be viewed with: [Microsoft PowerPoint Viewer](#)

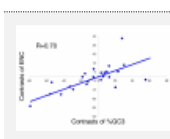


Figure 4. Correlation between phylogenetically independent contrasts of each species' %GC3 and its mean ENC for 25 nematode species. ENC, effective number of codons.

To ensure that differences in our available data for each species (for instance, cluster number and cluster length) were not creating artifacts in ENC values, quality checks were performed. Unlike measures such as codon bias index, scaled $\times 2$, and intrinsic codon bias index, ENC values should be independent of translated polypeptide length and sample size [49,51], and our analysis confirmed this. No correlation with ENC was observed with either average translated polypeptide length or number of clusters for a species. In fact, *SS* and *SR* with the lowest ENC values had above average cluster length and number. As additional confirmation, we randomly selected 2,400 *C. elegans* genes (the average number of clusters for species other than *CE* and *CB*) and calculated ENC based on either full-length genes or genes trimmed to 121 AAs (the average length cluster translation for species other than *CE* and *CB*). Differences in the average ENC numbers for these datasets were not statistically significantly different from zero ($P > 0.05$).

In addition to codon bias, neighboring nucleotides influence the codon observed at a position relative to synonymous codons. The most important nucleotide determining such context dependent codon bias [52-54] is the first one following the codon (N1 context) [55,56]. An analysis using the complete genesets of *Homo sapiens*, *Drosophila melanogaster*, *C. elegans*, and *Arabidopsis thaliana* revealed that 90% of codons have a statistically significant N1 context-dependent codon bias [57]. Using the same method we calculated that, for the 30 nematode species represented by EST-derived codon data, an average of 63% of codons with N1 context have a statistically significant bias (because the R values differed from 1 by more than 3 standard deviations). Fedorov and colleagues [57] showed that their results were not considerably affected by gene sampling. However, for our dataset the calculated *CE*-A and *CE*-B N1 context with statistically significant bias was 75% and 83% of the codons, respectively, as compared with 96% when the complete *C. elegans* gene set was used. Therefore, the extent of significant N1 context-dependent codon bias determined from EST-based codon usage data may change as more complete nematode genomes become available. The complete list of relative abundance of all nematode species with N1 context, R values, and standard deviations are available in Additional data file 4.

Additional file 4.

Format: XLS Size: 416KB [Download file](#)

This file can be viewed with: [Microsoft Excel Viewer](#)

Coding sequence GC content versus total genome GC content

Because of the clear relationships of AA composition, codon usage pattern, and codon bias to the GC content of coding sequences and the interest in the underlying cause of these correlations (see Discussion, below), we examined the relationship between coding sequence GC3 content and genomic GC content in nematodes. Total genomic GC content was calculated for the six nematode species for which significant genome sequence data were available as unassembled sequences (*TS* and *HC*), partial assemblies (*BM* and *AC*), or finished assemblies (*CE* and *CB*). Noncoding genomic GC content was calculated for *CB* and *CE* based on published estimates of the percentage of each genome that is composed of noncoding sequence, namely 74.5% and 77.1%, respectively [35]. Extrapolations were made for other species using the *CE* percentage noncoding estimate. Although GC content varies across the genome for some organisms (for example, isochors in vertebrates [58]), GC content is fairly uniform across the *C. elegans* genome [17]; furthermore, as yet there is no evidence of non-uniformity in other nematode genomes. A positive correlation was observed between coding GC3 content and both total GC content and extrapolated noncoding GC content ($R = 0.92$; Figure 5). Noncoding genomic sequences varied across a wider span of GC values than did coding sequences. In all six nematodes, coding sequences were somewhat more GC rich than were noncoding sequences (2-10%).

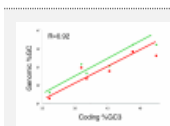


Figure 5. Correlation between coding sequence (transcriptome) %GC3 and genome %GC for six nematode species with extensive available genomic sequence. The green line indicates the coding sequence %GC versus the full genomic %GC. In this case, coding sequence %GC3 is a contributor to the full genomic %GC such that X and Y are not independent variables. The red line indicates the coding sequence %GC3 versus noncoding genomic %GC. In this case, the coding sequence contribution has been removed from genomic totals such that X and Y are independent variables. For *BM*, *TS*, *HC*, and *AC*, the calculation of noncoding genomic %GC relies on the assumption that the species will have a similar breakdown of coding and noncoding sequence as *CE*. Assembly and gene calling for the *BM*, *HC*, *TS*, and *AC* sequences are needed to test this assumption. Definitions of species two letter codes are provided in Table 1.

Comparison of coding sequence GC versus 3'-untranslated region (UTR) GC also supports this conclusion. Calculated 3'-UTR GC for the 30 species in our EST dataset ranges from 28.6% to 46.1%. Correlation between phylogenetically independent contrasts of coding GC content (Table 1) and 3'-UTR has an R value of 0.81 (data not shown).

Codon usage patterns in abundantly expressed genes and candidate optimal codons

Representation in cDNA library generally correlates with abundance in the original biologic sample [59] although artifacts occur [60,61]. To investigate the difference in the codon usage patterns in highly abundant transcripts as compared with less abundantly expressed genes, as determined by ESTs, we selected five species, each of which is a member of a different clade. The selected species (*AY*, *MI*, *OV*, *SR*, and *TS*) were represented by approximately 3,000 clusters each (range 2,693-3,214), and codon usage tables were generated for subsets of genes from each species: the 20 most abundant clusters versus all remaining clusters, and the 50 most abundant clusters versus all remaining clusters. Results of both comparisons were similar, and for simplicity we discuss only the results based on the comparison of the 50 most abundant versus all remaining clusters. Clusters 51 to about can be described as containing mainly genes with low to moderate expression because transcripts of extremely low abundance are less likely to be represented in EST collections (for instance, neuronal 7-transmembrane receptors). Codon usage tables, AA frequencies, and relative differences between AA usage of the most abundant and less abundant genes are available in Additional data file 5.

Additional file 5.

Format: XLS Size: 145KB [Download file](#)

D values were calculated across all AAs and the codon usage in each species was generally similar for genes represented by abundant EST clusters and genes represented by low to moderate expression EST clusters. *SR* exhibited the greatest difference between the two usage patterns ($D_{100} = 6.15$). Additionally, for all the species at least seven AAs were used significantly more frequently in the abundant genes than in the remainder of the genes. For example, although the abundant *OV* clusters had a Pro composition of 10.5% of all AAs, the rest of the clusters were only 4.4% Pro.

Examining the codon usage frequencies within an AA, an increase in usage has been noted with higher gene expression for specific so-called 'optimal' codons [62,63]. Using the codon usage tables for the top 50 and remaining clusters, we have defined a list of potentially optimal codons with usage that is higher in abundant transcripts by a statistically significant measure. Out of the 59 synonymous codons there were 24, 28, 25, 27, and 23 candidate optimal codons (Table 5) in *AY*, *MI*, *OV*, *SR*, and *TS*, respectively. For example, Tyr is encoded by two codons (TAC and TAT); in *AY* TAC is used 75% of the time in the abundant clusters and 59% of the time in the less abundant clusters. Similar analysis documented about 21 candidate optimal codons in *C. elegans* for which usage differed significantly when comparing high and low expressed genes [31,33,64]. Confirmation of these candidate codons as truly 'optimal' will require additional investigations, including other means of verifying relative expression levels (for example, microarrays and reverse transcription [RT]-polymerase chain reaction [PCR]).

Table 5. Candidate optimal codons in five species, determined as frequency increase by increased expression level^a

Discussion

A comprehensive and well supported codon usage table for 32 nematode species across most of the phylum's major clades and based on nearly 26 million codons is now available. Use of large EST datasets provide an excellent resource for determining a species mean codon usage with results that differ only modestly from those obtained from full genomes. In nematodes, codon usage varies widely, as does coding and noncoding GC content of nematode genomes. GC content correlates with AA usage, similarity of codon usage, and codon bias. Codon usage similarity in Nematoda usually persists within a genus but then rapidly diminishes, even within each major clade (clades I-V). Based on EST sampling, differences in codon usage between highly abundant genes and moderately expressed genes are recognizable, and candidate optimal codons can be identified.

GC content, causality, and directional mutation pressure

Correlations between GC content and mean codon usage and mean AA usage similar to those we describe across the phylum Nematoda have been observed in many other species [4,65-70]. Directional mutation pressure is a theory proposed to quantify differences in GC content observed in species [3]. Important variables include the relative values of the mutation rates u (GC/CG \rightarrow AT/TA change) and v (AT/TA \rightarrow GC/CG change). The preponderance of the evidence supports causality of genome GC content, as determined by directional mutation pressure or nucleotide level selective pressure, driving both codon usage and AA composition rather than the reverse. First, in an examination of sequence data from a large number of a bacteria, archaea, and eukaryotes, a model assuming directional mutation and selection at the nucleotide level with different rates of change for each of the three codon positions can explain 71-87% of the variance in codon usage and 71-79% of the variance in AA composition [5]. Knight and coworkers [5] found that between species an AA's change in frequency in response to GC content is determined by the mean GC content of its codons, whereas a codon's change in frequency is determined by the difference between its GC content and the mean GC content of its synonyms. We observe this result to be generally true across nematodes as well.

Second, an analysis comparing codon usage from eubacterial and archaeal species with complete genomes [6] found that codon usage can be predicted with some accuracy if one knows only the sequence of the species' intergenic sequences from which genome GC content, and context dependent nucleotide bias parameters can be calculated. Using data from six nematode species for which substantial genome sequence data are available, we observed that coding sequence GC3 content correlates with noncoding sequence GC content. This perhaps indicates that, for nematodes too, it should be possible to predict mean codon usage using only knowledge of the intergenic sequences of the species. Our findings are consistent with the model that genome GC content drives both mean codon usage and AA composition.

Little is known about why directional mutation pressure or selective pressure leads to differences in genomic GC content among species [5,6]. Within nematodes we see no pattern based on ecologic niche or other factors that corresponds to GC content. For instance, cyst nematodes (*GP*, *GR*, and *HG*) and root knot nematodes (*MI*, *MJ*, *MA*, *MH*, and *MC*) have similar life cycles as plant sedentary endoparasites, but their coding sequence GC contents are completely different (approximately 50% versus 36%). Whatever the driving forces, it is important for nematologists to note that they are sufficiently strong not only to change base composition in wobble sites (third position) but also to alter first and second codon positions and even AA sequences - features that are sometimes assumed to be under selective pressure for conservation.

Species' mean codon usage versus optimal codons in abundantly expressed genes

Our use of thousands of genes per species without weighting for abundance of expression has produced a codon usage dataset that probably reflects codon usage for genes with low to moderate abundance of mRNAs. In the case of *C. elegans* and *C. briggsae*, our codon usage table reflects the mean of all predicted genes, although this is similar to that observed based on sampling of 10,000 ESTs. At this 'genome-wide' level, genome GC content is a dominant factor. However, codon usage within a species does vary from gene to gene.

Prior studies of *C. elegans* codon usage have examined codon usage and the role of 'optimal codons' in putatively abundantly expressed genes [31,33,64]. Stenico and coworkers [31] observed differences between usage of specific codons based on 168 known genes, including many highly expressed transcripts (for instance, actin, myosin, collagen, and vitellogenin), and 90 unidentified reading frames (URFs) emerging from sequencing efforts presumed to represent a more 'random sampling' of the genome. Overall, our codon usage results based on the full *C. elegans* genome are similar to both the results from Stenico and coworkers' 168 known genes ($D_{100} = 0.97$) and the 90 URFs ($D_{100} = 1.1$). Duret and coworkers [33] weighted 15,425 *C. elegans* genes for expression levels based on their EST abundance and identified 21 favored codons used most frequently in highly expressed genes. In all cases, these optimal codons could be decoded by isoaccepting tRNAs that had the highest gene copy number in the genome, indicating that optimal codons are probably selected for translational efficiency. Likewise, Kanaya and coworkers [64] showed that, in *C. elegans*, ribosomal proteins and histones,

selected as representatives of highly expressed genes, also use optimal codons different from those used by average genes and that these optimal codons correspond to tRNA gene copy number. AA frequencies in abundant *C. elegans* genes also correspond to isoaccepting tRNA gene copy number ($R^2 = 0.67$) [33].

Therefore, in *C. elegans* different pictures emerge of evolutionary forces acting on codons and AAs in low to moderately expressed genes (directional mutation pressure, genome GC content) compared with abundantly expressed genes (optimal codons, tRNA copy number). In other nematodes, it is possible that a similar dichotomy exists, although we currently lack knowledge of tRNA gene copy number, and information on gene expression levels is largely limited to estimations based on EST abundance. Here, we have provided candidate optimal codons in *AY*, *MI*, *OV*, *SR*, and *TS*. A more detailed examination of codon usage as it relates to gene expression level in other nematodes will be possible by taking advantage of microarray and RF-PCR confirmation of transcript abundance.

Implications for phylogenetic studies and molecular biology

The extent to which average nematode genes sequences are susceptible to GC or AT shifts should sound a cautionary note for phylogenetic studies of nematode species, genes, and proteins based solely on coding sequences because convergent evolution may create confusing results. Knight and coworkers [5] noted that, 'Pairs of species with convergent GC contents might also evolve convergent protein sequences, especially at functionally unconstrained positions. For example, the frequencies of both lysine and arginine are highly (but oppositely) correlated with GC content, and lysine and arginine can easily substitute for one another in proteins.' In nematodes as well, one can envision exchanges of Lys and Arg (Figure 2).

For cloning genes of interest from various nematode species, we found that codon usage is a rapidly evolving feature such that codon usage patterns beyond within a genus comparisons are often divergent. Therefore, extrapolating assumed codon usage patterns to unsampled species in nematodes beyond the genus level is unlikely to be successful. At a practical level of species choice, cloning of orthologs and homologs of interest from species that are AT rich and have low ENC values, such as *SS* and *MI* with low ENC values, will require fewer degenerate primers than may be needed for more GC rich species such as *TC* and *MI*. Transcript abundance is also an important factor because genes suspected of high level expression are likely to exhibit a shift in their codon usage from the species average toward optimal codons selected for translational efficiency.

Conclusion

Extensive sequence datasets from one complete, one draft, and 30 partial genomes across the phylum Nematoda have been used to analyze the conservation and diversification of encoded protein families [26] and the factors effecting codon usage and bias (the present report). The undertaken comprehensive survey of observed codon usage and bias is based on 26 million codons in 32 species, making it the most extensive study for any phylum. Our data indicate that similarity between species in average codon usage is a short range phenomenon, generally rapidly diminishing beyond the genus level. Mapping codon usage changes to the phyla indicates the genera *Globodera*, *Meloidogyne*, *Pristionchus*, and *Strongyloides* have the most highly derived patterns of codon usage in nematodes, with the remaining species exhibiting less relatively divergence from an ancestral nematode pattern. There was a strong correlation between the exonic GC content and similarity in codon usage. GC content also explains much of the observed variation in the effective number of codons, a measure of codon bias, and even differences in AA frequency. Results from partial genomes assembled from ESTs and complete genomes provide generally good agreement on codon usage, although refinement will be necessary as more sequences become available. EST collections from five species have also been used as a starting point to identify potentially abundant genes and predict optimal codons. These predictions will also be refined using more accurate measures of gene expression, including microarrays and quantitative RT-PCR.

Materials and methods

Sequence acquisition and organization

To perform the first meta-analysis of the genomic biology of the phylum Nematoda [26], ESTs from 30 nematode species generated by our laboratories and others were downloaded from the dbEST division of GenBank in May 2003. For consistency, in this accompanying analysis of codon usage we used this dataset for all analyses. Sequences were collated and processed into partial genomes using the PartiGene pipeline [71,72]. Polypeptide translations were predicted using prot4EST [45,72]. Wormpep_dna121 (March 2004; Wellcome Trust Sanger Institute, unpublished data) was used for *C. elegans* analysis, and the hybrid gene set [35] was used for *C. briggsae* analysis. Mitochondria can have codon usage differing from that of the nuclear genome, and therefore protein coding genes from mitochondrial genomes were eliminated from consideration. Codon usage tables for human, *Saccharomyces cerevisiae*, and *Escherichia coli* were derived from the Codon Usage Table Database [73] derived from GenBank Release 140.0 (22 March 2004 [74]).

Phylogenetic correction

Analyses of the relationships among GC content, AA, and codon usage values require statistical correction for the phylogenetic relatedness of the species being studied using phylogenetically independent contrasts [75]. To generate these contrasts, we performed the following procedures. First, to construct a phylogenetic tree independent of the transcriptomic data analyzed in this paper, we aligned 18S ribosomal RNA sequences using Clustalw [76] for all nematodes for which more than 15 kilobases of sequence was available. The 18S sequence GenBank accession numbers are available in Figure 1; the sequences from a priapulid worm and a nematomorph were used as outgroups [16] but excluded from our analysis. Alignments were trimmed to reflect only the overlapping portion of the sequences from all species analyzed. Second, this alignment, containing 1,841 base pairs/species (including gaps) and an alternative alignment excluding any region involving an insertion or deletion (1,423 base pairs/species remained), was used to estimate phylogenies from the nucleotide sequences by parsimony and maximum likelihood (with and without assumption of a molecular clock) using Phylip [77]. Third, the trees with branch lengths derived from molecular clock-based analysis were used to calculate phylogenetically independent contrasts for our parameters of interest [75]. The Phylip program 'contrasts' was used to compute the phylogenetically independent contrasts using a Brownian-motion model [78,79] of genomic evolution.

Bioinformatics

The Emboss program 'cusp' was used to calculate codon usage in the predicted translations [80]. The ENC [49] was calculated using the Emboss program 'Codon Heterozygosity (Inverse of) in Protein-coding Sequences'. A genetic distance statistic was used to quantify divergence of synonymous codon usage between species [47] follows. Let t_j be the number of codons that code for the j^{th} amino acid. We omit analysis of the nondegenerate codons Met and Trp, as well as the 'stop' codon, so that $j = 1, 2 \dots r$, where $r = 18$. Furthermore,

let a_{ij} and b_{ij} be the i^{th} synonymous codon in the j^{th} AA of two organisms A and B, respectively. Then, Nei's difference statistic D is defined as the following:

$$J_{jaa} = \sum_{i=1}^{t_j} a_{ij}^2 \quad J_{jbb} = \sum_{i=1}^{t_j} b_{ij}^2 \quad J_{jab} = \sum_{i=1}^{t_j} a_{ij}b_{ij}$$
$$J_{aa} = \frac{1}{r} \sum_{j=1}^r J_{jaa} \quad J_{bb} = \frac{1}{r} \sum_{j=1}^r J_{jbb} \quad J_{ab} = \frac{1}{r} \sum_{j=1}^r J_{jab}$$
$$D = -\log\left(\frac{J_{ab}}{\sqrt{J_{aa}J_{bb}}}\right).$$

Investigators have used D as an empirical measure of difference, averaged over all r residues, of the codon usage between organisms [48]. There are a total of $C_{32,2} = 496$ meaningful comparisons for the entire collection of 32 species. These results are presented as an $N \times N$ square matrix and the values are presented as $D \times 100$. For simplicity in the remainder of the text we will refer to $D \times 100$ as D_{100} .

Phylogenetic changes in codon usage were analyzed using the species tree derived from 18S rRNA sequences estimated by maximum likelihood with a molecular clock imposed. Partitioning a matrix of distance values on a phylogenetic tree can estimate amounts of change occurring on each branch, provided that the distance metrics obey the triangle inequality (see the discussion on page 25 of the report by Page and Holmes [81]). Because of its logarithmic operation, Nei's difference statistic D violates the triangle inequality at high values. For the phylogenetic analysis of codon usage, we substituted for D a distance measure equal to $1 - \text{antilog}(-D)$, which obeys the triangle inequality. Distances were partitioned on the tree topology using the cyclic neighbor-joining algorithm illustrated by Avise [82], except that the topology was specified by the prior analysis of 18S rRNA sequences.

The ENC was used to measure the degree of codon bias for a gene [49]. Because the ENC statistic is not reliable when analyzing very short sequences (20 AAs or less), 54 short translations out of a total of 70,358 were discarded from these analyses. The relative abundance of nematode codons (per species) having a statistically significant N1 context-dependent codon bias was calculated by computing the R values and the standard deviations, as described by Fedorov and coworkers [57].

Predicted expression level of a transcript (abundant, moderate, rare, and so on) was determined by counting the number of ESTs comprising the cluster. Five species from different clades that have been sampled with at least 10,000 ESTs from several life-stage libraries were selected for these analyses. Most of the cDNA libraries were constructed using the same protocols [61,83], and although the libraries generally correlate with abundance in the original biologic sample, artifacts can occur. The increase in use of a given codon for an AA in highly expressed genes (optimal codons) was considered significant when the difference of the codon distributions within that AA was statistically significant between datasets ($P \leq 0.01$).

To assess the differences in calculated codon usage distributions when using partial (EST-based) as compared with whole genome data, we generated two datasets using *C. elegans* ESTs and compared them with the curated gene set of *C. elegans* (Wormpep version 121). Each EST dataset was composed of 10,000 ESTs (approximately the average number of ESTs used for the other 30 species); clustering and peptide predictions were performed using the same algorithms as for the other species.

Additional data files

The following additional data are included with the online version of this article: An Excel file containing a table that shows the nucleotide usage (%) by codon position and nematode species (Additional data file 1); an Excel file containing an $N \times N$ square matrix that shows codon usage across degenerate AAs for 25 nematode species reported as D values (Additional data file 2); a PowerPoint file containing a figure that shows the distribution of genes with various degrees of codon usage bias as measured by ENC for three species with approximately the same number of clusters but with different coding GC content (*S. ratti* [GC = 32%], *P. trichosuri* [40%], and *P. pacificus* [51%]; Additional data file 3); an Excel file containing a table that shows the N1 context dependent bias per species (Additional data file 4); and an Excel file containing a table that shows codon usage of abundant and less abundant translations for five nematode species (Additional data file 5).

Acknowledgements

We are thankful for assistance from Barry Shortt with statistics, Matt Dimmic with informatics, and Mark Blaxter and LaDeana Hillier with useful comments on the manuscript. Work at Washington University was supported by NIH-NIAID research grant AI 46593 to RHW and Richard K Wilson. JP was supported in part by the Wellcome Trust. The project originated while JPM was a Merck Fellow of the Helen Hay Whitney Foundation. JPM is an employee and equity holder of Divergence, Inc.

References

1. Milhon JL, Tracy JW: **Updated codon usage in *Schistosoma*.** *Exp Parasitol* 1995, **80**:353-356. [PubMed Abstract](#) | [Publisher Full Text](#)
2. Duret L: **Evolution of synonymous codon usage in metazoans.** *Curr Opin Genet Dev* 2002, **12**:640-649. [PubMed Abstract](#) | [Publisher Full Text](#)
3. Sueoka N: **On the genetic basis of variation and heterogeneity of DNA base composition.** *Proc Natl Acad Sci USA* 1962, **48**:582-592. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
4. Sueoka N: **Directional mutation pressure and neutral molecular evolution.** *Proc Natl Acad Sci USA* 1988, **85**:2653-2657. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
5. Knight RD, Freeland SJ, Landweber LF: **A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes.** *Genome Biol* 2001, **2**:research0010. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)

6. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH: **Codon usage between genomes is constrained by genome-wide mutational processes.**
Proc Natl Acad Sci USA 2004, **101**:3480-3485. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
7. Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.**
Mol Biol Evol 1985, **2**:13-34. [PubMed Abstract](#) | [Publisher Full Text](#)
8. Moriyama EN, Powell JR: **Codon usage bias and tRNA abundance in *Drosophila*.**
J Mol Evol 1997, **45**:514-523. [PubMed Abstract](#) | [Publisher Full Text](#)
9. Bulmer M: **Coevolution of codon usage and transfer RNA abundance.**
Nature 1987, **325**:728-730. [PubMed Abstract](#) | [Publisher Full Text](#)
10. Sharp PM, Li WH: **The codon Adaptation Index: a measure of directional synonymous codon usage bias, and its potential applications.**
Nucleic Acids Res 1987, **15**:1281-1295. [PubMed Abstract](#) | [PubMed Central Full Text](#)
11. Gouy M, Gautier C: **Codon usage in bacteria: correlation with gene expressivity.**
Nucleic Acids Res 1982, **10**:7055-7074. [PubMed Abstract](#) | [PubMed Central Full Text](#)
12. Carlini DB, Chen Y, Stephan W: **The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*.**
Genetics 2001, **159**:623-633. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
13. Chamary JV, Hurst LD: **Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.**
Genome Biol 2005, **6**:R75. [PubMed Abstract](#) | [BioMed Central Full Text](#) | [PubMed Central Full Text](#)
14. Oresic M, Dehn M, Korenblum D, Shalloway D: **Tracing specific synonymous codon-secondary structure correlations through evolution.**
J Mol Evol 2003, **56**:473-484. [PubMed Abstract](#) | [Publisher Full Text](#)
15. Marquez R, Smit S, Knight R: **Do universal codon-usage patterns minimize the effects of mutation and translation error?**
Genome Biol 2005, **6**:R91. [PubMed Abstract](#) | [BioMed Central Full Text](#) | [PubMed Central Full Text](#)
16. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, *et al.*: **A molecular evolutionary framework for the phylum Nematoda.**
Nature 1998, **392**:71-75. [PubMed Abstract](#) | [Publisher Full Text](#)
17. The *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans* : a platform for investigating biology.**
Science 1998, **282**:2012-2018. [PubMed Abstract](#) | [Publisher Full Text](#)
18. Blaxter ML, Raghavan N, Ghosh I, Guiliano D, Lu W, Williams SA, Slatko B, Scott AL: **Genes expressed in *Brugia malayi* infective third stage larvae.**
Mol Biochem Parasitol 1996, **77**:77-93. [PubMed Abstract](#) | [Publisher Full Text](#)
19. Dautova M, Rosso MN, Abad P, Gommers FJ, Bakker J, Smart G: **Single pass cDNA sequencing - a powerful tool to analyse gene expression in preparasitic juveniles of the southern root-knot nematode *Meloidogyne incognita*.**
Nematology 2001, **3**:129-139. [Publisher Full Text](#)
20. McCarter J, Dautova Mitreva M, Martin J, Dante M, Wylie T, Rao U, Pape D, Bowers Y, Theising B, Murphy CV, *et al.*: **Analysis and functional classification of transcripts from the Nematode *Meloidogyne incognita*.**
Genome Biol 2003, **4**:R26. [PubMed Abstract](#) | [BioMed Central Full Text](#) | [PubMed Central Full Text](#)
21. Daub J, Loukas A, Pritchard DI, Blaxter M: **A survey of genes expressed in adults of the human hookworm, *Necator americanus*.**
Parasitology 2000, **120**:171-184. [PubMed Abstract](#) | [Publisher Full Text](#)
22. Mitreva M, McCarter JP, Martin J, Dante M, Wylie T, Chiapelli B, Pape D, Clifton SW, Nutman TB, Waterston RH: **Comparative genomics of gene expression in the parasitic and free-living nematodes *Strongyloides stercoralis* and *Caenorhabditis elegans*.**
Genome Res 2004, **14**:209-220. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
23. Thompson FJ, Mitreva M, Barker GL, Martin J, Waterston RH, McCarter JP, Viney ME: **An expressed sequence tag analysis of the life-cycle of the parasitic nematode *Strongyloides ratti*.**
Mol Biochem Parasitol 2005, **142**:32-46. [PubMed Abstract](#) | [Publisher Full Text](#)
24. Parkinson J, Mitreva M, Hall N, Blaxter M, McCarter JP: **400000 nematode ESTs on the Net.**
Trends Parasitol 2003, **19**:283-286. [PubMed Abstract](#) | [Publisher Full Text](#)
25. Mitreva M, Blaxter ML, Bird DM, McCarter JP: **Comparative genomics of nematodes.**
Trends Genet 2005, **21**:573-581. [PubMed Abstract](#) | [Publisher Full Text](#)
26. Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Hall N, Barrell B, Waterston RH, McCarter JP, *et al.*: **A transcriptomic analysis of the phylum Nematoda.**
Nat Genet 2004, **36**:1259-1267. [PubMed Abstract](#) | [Publisher Full Text](#)
27. **Nematode Net** [<http://www.nematode.net/>] [webcite](#)

28. Wylie T, Martin J, Dante M, Mitreva M, Clifton SW, Chinwalla A, Waterston RH, Wilson RK, McCarter JP: **Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes.**
Nucleic Acids Res 2004, **32**:D423-D426. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
29. **Nembase** [<http://www.nematodes.org/>] [webcite](#)
30. Parkinson J, Whitton C, Schmid R, Thomson M, Blaxter M: **NEMBASE: a resource for parasitic nematode ESTs.**
Nucleic Acids Res 2004, **32**:D427-D430. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
31. Stenico M, Lloyd AT, Sharp PM: **Codon usage in *Caenorhabditis elegans* : delineation of translational selection and mutational biases.**
Nucleic Acids Res 1994, **22**:2437-2446. [PubMed Abstract](#) | [PubMed Central Full Text](#)
32. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*.**
Proc Natl Acad Sci USA 1999, **96**:4482-4487. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
33. Duret L: **tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes.**
Trends Genet 2000, **16**:287-289. [PubMed Abstract](#) | [Publisher Full Text](#)
34. Marais G, Duret L: **Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*.**
J Mol Evol 2001, **52**:275-280. [PubMed Abstract](#) | [Publisher Full Text](#)
35. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, et al.: **The genome sequence of *Caenorhabditis briggsae* : a platform for comparative genomics.**
PLoS Biol 2003, **1**:E45. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
36. Unnasch TR, Katholi CR, Coate LM: ***Onchocerca volvulus* : frequency of codon usage.**
Exp Parasitol 1992, **75**:457-459. [PubMed Abstract](#) | [Publisher Full Text](#)
37. Hammond MP: **Codon usage and gene organization in *Brugia*.**
Parasitol Res 1994, **80**:173-175. [PubMed Abstract](#) | [Publisher Full Text](#)
38. Ellis J, Morrison DA, Kalinna B: **Comparison of the patterns of codon usage and bias between *Brugia*, *Echinococcus*, *Onchocerca* and *Schistosoma* species.**
Parasitol Res 1995, **81**:388-393. [PubMed Abstract](#) | [Publisher Full Text](#)
39. Fadiel A, Lithwick S, Wanas MQ, Cuticchia AJ: **Influence of intercodon and base frequencies on codon usage in filarial parasites.**
Genomics 2001, **74**:197-210. [PubMed Abstract](#) | [Publisher Full Text](#)
40. Moore TA, Ramachandran S, Gam AA, Neva FA, Lu W, Saunders L, Williams SA, Nutman TB: **Identification of novel sequences and codon usage in *Strongyloides stercoralis*.**
Mol Biochem Parasitol 1996, **79**:243-248. [PubMed Abstract](#) | [Publisher Full Text](#)
41. Fadiel AA, Lithwick S, Gamra MM: **Codon usage analysis of *Ascaris* species influence of base and intercodon frequencies on the synonymous codon usage.**
J Egypt Soc Parasitol 2002, **32**:625-638. [PubMed Abstract](#)
42. Fadiel AA, Lithwick S, el-Garhy MF: **Influence of parasitic life style on the patterns of codon usage and base frequencies of *Ancylostoma* and *Necator* species.**
J Egypt Soc Parasitol 2002, **32**:657-673. [PubMed Abstract](#)
43. Nakamura Y, Gojobori T, Ikemura T: **Codon usage tabulated from the international DNA sequence databases; its status 1999.**
Nucleic Acids Res 1999, **27**:292. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
44. McCarter JP, Bird DM, Mitreva M: **Nematode gene sequences: update for December 2005.**
J Nematol 2006, **37**:417-421.
45. Wasmuth JD, Blaxter ML: **prot4EST: translating expressed sequence tags from neglected genomes.**
BMC Bioinformatics 2004, **5**:187. [PubMed Abstract](#) | [BioMed Central Full Text](#) | [PubMed Central Full Text](#)
46. King JL, Jukes TH: **Non-Darwinian evolution.**
Science 1969, **164**:788-798. [PubMed Abstract](#)
47. Nei M: **Genetic distance between populations.**
Am Naturalist 1972, **106**:283-292. [Publisher Full Text](#)
48. Long M, Gillespie JH: **Codon usage divergence of homologous vertebrate genes and codon usage clock.**
J Mol Evol 1991, **32**:6-15. [PubMed Abstract](#) | [Publisher Full Text](#)
49. Wright F: **The 'effective number of codons' used in a gene.**
Gene 1990, **87**:23-29. [PubMed Abstract](#) | [Publisher Full Text](#)
50. Powell JR, Moriyama EN: **Evolution of codon usage bias in *Drosophila*.**

51. Comeron JM, Aguade M: **An evaluation of measures of synonymous codon usage bias.**
J Mol Evol 1998, **47**:268-274. [PubMed Abstract](#) | [Publisher Full Text](#)
52. Yarus M, Folley LS: **Sense codons are found in specific contexts.**
J Mol Biol 1984, **182**:529-540. [Publisher Full Text](#)
53. Shpaer EG: **Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation.**
J Mol Biol 1986, **188**:555-564. [PubMed Abstract](#) | [Publisher Full Text](#)
54. Gouy M: **Codon contexts in *Enterobacterialand coliphage* genes.**
Mol Biol Evol 1987, **4**:426-444. [PubMed Abstract](#) | [Publisher Full Text](#)
55. Berg OG, Silva PJN: **Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection.**
Nucleic Acids Res 1997, **25**:1397-1404. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
56. Karlin S, Mrazek J: **What drives codon choices in human genes?**
J Mol Biol 1996, **262**:459-472. [PubMed Abstract](#) | [Publisher Full Text](#)
57. Fedorov A, Saxonov S, Gilbert W: **Regularities of context-dependent codon bias in eukaryotic genes.**
Nucleic Acids Res 2002, **30**:1192-1197. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
58. Eyre-Walker A, Hurst LD: **The evolution of isochores.**
Nat Rev Genet 2001, **2**:549-555. [PubMed Abstract](#) | [Publisher Full Text](#)
59. Audic S, Claverie JM: **The significance of digital gene expression profiles.**
Genome Res 1997, **7**:986-995. [PubMed Abstract](#) | [Publisher Full Text](#)
60. Munoz ET, Bogarad LD, Deem MW: **Microarray and EST database estimates of mRNA expression levels differ: the protein length versus expression curve for *C. elegans*.**
BMC Genomics 2004, **5**:30. [PubMed Abstract](#) | [BioMed Central Full Text](#) | [PubMed Central Full Text](#)
61. Mitreva M, Jasmer DP, Appleton J, Martin J, Dante M, Wylie T, Clifton SW, Waterston RH, McCarter JP: **Gene discovery in the adenophorean nematode *Trichinella spiralis*: an analysis of transcription from three life cycle stages.**
Mol Biochem Parasitol 2004, **137**:277-291. [PubMed Abstract](#) | [Publisher Full Text](#)
62. Ikemura T: **Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system.**
J Mol Biol 1981, **151**:389-409. [PubMed Abstract](#) | [Publisher Full Text](#)
63. Musto H, Romero H, Zavala A, Jabbari K, Bernardi G: **Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection.**
J Mol Evol 1999, **49**:27-35. [PubMed Abstract](#) | [Publisher Full Text](#)
64. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T: **Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis.**
J Mol Evol 2001, **53**:290-298. [PubMed Abstract](#) | [Publisher Full Text](#)
65. Ohama T, Yamao F, Muto A, Osawa S: **Organization and codon usage of the streptomycin operon in *Micrococcus luteus*, a bacterium with a high genomic G + C content.**
J Bacteriol 1987, **169**:4770-4777. [PubMed Abstract](#) | [PubMed Central Full Text](#)
66. Ohama T, Muto A, Osawa S: **Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content.**
Nucleic Acids Res 1990, **18**:1565-1569. [PubMed Abstract](#) | [PubMed Central Full Text](#)
67. Sueoka N: **Directional mutation pressure, selective constraints, and genetic equilibria.**
J Mol Evol 1992, **35**:95-114.
68. Wilquet V, Van de Castele M: **The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition.**
Res Microbiol 1999, **150**:21-32. [PubMed Abstract](#) | [Publisher Full Text](#)
69. D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G: **Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins.**
J Mol Evol 1991, **32**:504-510. [PubMed Abstract](#) | [Publisher Full Text](#)
70. Lobry JR: **Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species.**
Gene 1997, **205**:309-316. [PubMed Abstract](#) | [Publisher Full Text](#)
71. Parkinson J, Guiliano DB, Blaxter M: **Making sense of EST sequences by CLOBBing them.**
BMC Bioinformatics 2002, **3**:31. [PubMed Abstract](#) | [BioMed Central Full Text](#) | [PubMed Central Full Text](#)
72. Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M: **PartiGene: constructing partial genomes.**
Bioinformatics 2004, **20**:1398-1404. [PubMed Abstract](#) | [Publisher Full Text](#)
73. **Codon Usage Table Database** [<http://www.kazusa.or.jp/codon/>] [webcite](#)

74. Nakamura Y, Gojobori T, Ikemura T: **Codon usage tabulated from international DNA sequence databases: status for the year 2000.**
Nucleic Acids Res 2000, **28**:292. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
75. Felsenstein J: **Phylogenies and the comparative method.**
Am Naturalist 1985, **125**:1-12. [Publisher Full Text](#)
76. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.**
Nucleic Acids Res 2003, **31**:3497-3500. [PubMed Abstract](#) | [Publisher Full Text](#) | [PubMed Central Full Text](#)
77. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).**
Cladistics 1989, **5**:164-166.
78. Cavalli-Sforza LL, Edwards AWF: **Phylogenetic analysis: models and estimation procedures.**
Evolution 1967, **32**:550-570. [Publisher Full Text](#)
79. Edwards AWF, Cavalli-Sforza LL: **Reconstruction of evolutionary trees.** In *Phenetic and Phylogenetic Classification. Volume 6.* Edited by: Heywood VH, McNeill. London: Systematics Association; 1964:67-76.
80. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.**
Trends Genet 2000, **16**:276-277. [PubMed Abstract](#) | [Publisher Full Text](#)
81. Page RDM, Holmes EC: *Molecular Evolution: a Phylogenetic Approach.* Oxford, UK: Blackwell Science; 1998.
82. Avise JC: *Molecular Markers Natural History and Evolution.* New York: Chapman and Hall; 1994.
83. Mitreva M, Elling AA, Dante M, Kloek AP, Kalyanaraman A, Aluru S, Clifton SW, Bird DM, Baum TJ, McCarter JP: **A survey of SL1-spliced transcripts from the root-lesion nematode *Pratylenchus penetrans*.**
Mol Gen Genomics 2004, **272**:138-148. [Publisher Full Text](#)

[Have something to say? Post a comment on this article!](#)

