The Evolution of Biased Codon and Amino Acid Usage in Nematode Genomes

Asher D. Cutter,[1] *James D. Wasmuth,*[2] *and Mark L. Blaxter*

Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom

Despite the degeneracy of the genetic code, whereby different codons encode the same amino acid, alternative codons and amino acids are utilized nonrandomly within and between genomes. Such biases in codon and amino acid usage have been demonstrated extensively in prokaryote genomes and likely reflect a balance between the action of mutation, selection, and genetic drift. Here, we quantify the effects of selection and mutation drift as causes of codon and amino acid–usage bias in a large collection of nematode partial genomes from 37 species spanning approximately 700 Myr of evolution, as inferred from expressed sequence tag (EST) measures of gene expression and from base composition variation. Average G + C content at silent sites among these taxa ranges from 10% to 63%, and EST counts range more than 100-fold, underlying marked differences between the identities of major codons and optimal codons for a given species as well as influencing patterns of amino acid abundance among taxa. Few species in our sample demonstrate a dominant role of selection in shaping intragenomic codon-usage biases, and these are principally free living rather than parasitic nematodes. This suggests that deviations in effective population size among species, with small effective sizes among parasites, are partly responsible for species differences in the extent to which selection shapes patterns of codon usage. Nevertheless, a consensus set of optimal codons emerges that is common to most taxa, indicating that, with some notable exceptions, selection for translational efficiency and accuracy favors similar sets of codons regardless of the major codon-usage trends defined by base compositional properties of individual nematode genomes.

Introduction

The degeneracy of the genetic code allows for multiple codons to encode the same amino acid. However, degenerate codons are not present at equal frequencies in genes, a phenomenon termed codon-usage bias (Grantham et al. 1980; Sharp et al. 1995; Duret 2002). Codon-usage bias can be driven by the neutral processes of mutation, genetic drift, and/or biased gene conversion, so the relative abundance of alternative codons might reflect skews in local base composition (Sueoka 1988; Marais 2003). Additionally, selection for translational efficiency and/or accuracy can skew codon frequencies toward "optimal" codons (Ikemura 1982; Duret 2002). Selection on codon usage can be inferred from genomic correlations with the relative abundance of alternative tRNA molecules or gene copies, gene expression levels, synonymous substitution rates, or skewed levels of polymorphism at synonymous sites (Bennetzen and Hall 1982; Sharp and Li 1987; Akashi 1995; Duret and Mouchiroud 1999)—although an ongoing problem is to quantify the relative importance of selective and neutral forces as causes of codon-usage bias within and between species.

Because the fitness differences associated with the usage of alternative codons are subtle, the selection coefficients (s) involved in adaptive codon-usage bias are very small ($s \sim 10^{-6}$), thus requiring large effective population sizes (N_e) to offset the stochastic effects of genetic drift ($N_e \sim s^{-1}$) (Li 1987; Bulmer 1991; Akashi 1995). Indeed, genomes exhibiting the strongest biases in codon usage correspond to species of bacteria and yeast, which can have effective population sizes greatly in excess of 10^6 (Ikemura 1982; Merkl 2003). The genomes of *Drosophila* species also have extensive codon-usage bias, as do species of *Caenorhabditis* and *Arabidopsis* (Stenico et al. 1994; Akashi 1995; Kreitman and Antezana 1999; Wright et al. 2004). Despite skewed codon usage in mammals, natural selection does not appear to play a role (Ikemura 1985; Urrutia and Hurst 2001), with the possible exception of exonic regions involved in splicing (Parmley et al. 2006). General differences in patterns of codon usage between species are thought principally to be due to mutational processes on base composition (Knight et al. 2001; Chen et al. 2004). Brownian motion models may capture the predominant dynamics in the divergence of genomic base composition (Haywood-Farmer and Otto 2003) and, therefore, may also describe interspecific dynamics of overall codon-usage trends. However, intraspecific variation fits neutral mutational models less well, suggesting that deviations in the effectiveness of selection among loci is likely an important force shaping patterns of intragenomic codon-usage variation across all domains of life (Knight et al. 2001).

In addition to changes in overall trends in codon usage, species can evolve different optimal codons for a given amino acid. Changes in optimal codon identity will be difficult to achieve in genomes subject to consistent selection favoring particular alternative codons because 1) a change in optimal codon identity will result in substantial genetic load, due to the immediate selective costs of those highly expressed genes that contain high frequencies of the prior optimal codon (which is now nonoptimal) and 2) such shifts likely require alterations in tRNA gene abundances in a genome. Thus, evolutionary transitions in the identity of optimal codons are expected to occur only rarely, although this issue has received relatively little attention (Kreitman and Antezana 1999; McVean and Vieira 1999; Herbeck and Novembre 2003; Wall and Herbeck 2003). Shifts in the identity of optimal codons may be facilitated by a period of relaxed selection on codon usage (due to reduced effective population size), permitting changes in isoaccepting tRNA gene abundance and codon frequencies to accumulate by mutation drift, so that subsequent, more effective selection (through increased effective population size) could yield different optimal codons. Although genomic

[1] Present address: Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada.

[2] Present address: Department of Genetics and Genomic Biology, Hospital for Sick Children, Toronto, Ontario, Canada.

Key words: codon-usage bias, translational selection, molecular evolution, *Caenorhabditis elegans*.

E-mail: asher.cutter@utoronto.ca.

Mol. Biol. Evol. 23(12):2303–2315. 2006
doi:10.1093/molbev/msl097
Advance Access publication August 25, 2006

© The Author 2006. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

analyses of codon bias have provided robust descriptions for prokaryote and individual eukaryote genomes, the few taxonomically dense studies available in eukaryotes focus on individual genes (Morton and Levin 1997; Herbeck and Novembre 2003; Wall and Herbeck 2003). A more complete comparative context requires simultaneous analysis of codon bias for collections of many genes from many eukaryote taxa.

Processes that shape nonrandom usage of alternative codons also have the potential to skew the relative abundance of different amino acids used in proteins. This can occur due to neutral processes because the base compositions of all the codons encoding a given amino acid may be GC rich or GC poor (Foster et al. 1997). Alternatively, selection may skew amino acid frequencies because functionally similar amino acids may have different tRNA abundances or require different metabolic costs for their production (Barrai et al. 1995; Akashi and Gojobori 2002; Seligmann 2003). Base composition in a number of species has been shown to correlate with the amino acid content of proteins (Sueoka 1961; D'Onofrio et al. 1991; Foster et al. 1997; Lobry 1997; Gu et al. 1998; Singer and Hickey 2000); likewise, abundant and rare proteins can have different amino acid profiles (Akashi and Gojobori 2002; Merkl 2003). However, gene function may confound the interpretation of differences in amino acid frequencies of the encoded proteins; for example, highly abundant proteins might share similar functions, so similarity in amino acid profiles among them could simply reflect their common peptide domains rather than selection for efficient and/or accurate translation.

Here, we characterize patterns of codon-usage bias for partial genomes of 37 nematode species, using a large sample of expressed sequence tags (ESTs; 248,000 plus 257,000 from *Caenorhabditis elegans*) corresponding to nearly 100,000 genes (Parkinson, Mitreva, et al. 2004). We infer the set of optimal codons for each species and describe the relative importance of neutral and selective forces in shaping skews in the usage of degenerate codons and different amino acids. We find that selection on codon usage is widespread in free-living nematode species and, correspondingly, that these species or their recent ancestors are likely to have very large effective population sizes. However, most of the parasitic species show little evidence for selection dominating their biases in codon usage. We suggest that the parasitic lifestyle limits their effective population sizes and, therefore, that the stochastic processes of mutation and genetic drift largely determine their patterns of skew in codon usage.

Materials and Methods

EST Inference

The collection of ESTs for each species derives from a collaborative sequencing effort for a large number of nematode species (Parkinson, Mitreva, et al. 2004; Mitreva et al. 2005). For brevity, we refer to the 36 species included in this study by their 2-letter designations indicated in table 1. All ESTs from these species were processed with the PartiGene system, an integrated sequence analysis suite for transcriptomic data (Parkinson, Anthony, et al. 2004). To reduce the redundancy of the EST data sets, the sequences were first clustered using the CLOBB program (Parkinson et al. 2002), and consensus sequences for each cluster were assembled with Phrap (Ewing and Green 1998). Because ESTs derive from single-pass reads, most ESTs cover only part of the transcribed mRNA and may have base-call errors including reading frameshifts or ambiguous bases. Furthermore, an EST may be composed partly or completely of untranslated region and, therefore, not represent any part of the polypeptide sequence. To overcome these obstacles for generating EST consensus clusters for inferring correct coding sequence, we implemented prot4EST for peptide translation (Wasmuth and Blaxter 2004). prot4EST compares the peptide predictions of several translation algorithms and retrieves the most plausible translation. The parameters for prot4EST were optimized separately for each nematode species, collectively yielding the nematode peptide database NemPep (JD Wasmuth, unpublished data). NemPep v. 3 (June 2005) was used for these analyses, with the EST clusters and their polypeptide translations available through NEMBASE (Parkinson, Whitton, et al. 2004). Data labeled as *Parastrongyloides trichosuri* sequences in NEMBASE were not included in the analysis because we identified a strongly bimodal distribution of G + C at 4-fold silent sites (modes at $\sim$12% and $\sim$50%), raising doubts about the species integrity of this data set. Hereafter, we refer to the 116,919 EST clusters derived from 314,095 ESTs and their peptide translations used in this analysis simply as "genes," recognizing that in most cases they do not represent full-length coding sequences. ESTs predicted to correspond to mitochondrial genes were excluded from analysis, and all analyses were limited to the subset of 82,677 genes with $\geq$100 codons. For comparison, we also acquired 14,527 *C. elegans* full-length coding sequences that had corresponding ESTs available from Wormbase release WS140 (257,027 ESTs total; only one splice form per gene was considered).

Codon- and Amino Acid–Usage Calculations and Analysis

For each gene, we computed codon-usage bias with ENC, the effective number of codons (Wright 1990), and F_{op}, the frequency of optimal codons inferred from ΔRSCU analysis (Ikemura 1985; Duret and Mouchiroud 1999)(see below). ENC, calculated here with the program INCA v2.0 (Supek and Vlahovicek 2004), measures departures from uniform codon usage without dependence on sequence length or specific knowledge of preferred codons, although it is affected by base composition (Comeron and Aguade 1998; Novembre 2002). A variant of ENC, N'_c, was also calculated with INCA in an attempt to take account of background base composition by using average nucleotide frequencies among ESTs for a given species (Novembre 2002); however, the lack of direct ortholog comparisons and of noncoding sequence information for these ESTs limits the potential advantages of the N'_c statistic. After inferring optimal codons, we calculated F_{op} using codonW with customized optimal codon tables (J Peden, http://codonw.sourceforge.net). We also computed the relative synonymous codon usage (RSCU) of each codon in each gene, which quantifies the abundance

Table 1
Summary of Species Included in Analysis

ID	Species	Clade[a]	Host, Reproduction, Transmission[b]	Number of Genes[c]	Mean GC3s	Mean ENC	Mean N_c'	Mean F_{op}	$\overline{\Delta RSCU_+}$
AC	*Ancylostoma caninum*	V	Canine, G, D	2,814	0.429	55.0	55.2	0.388	0.184
AY	*Ancylostoma ceylanicum*	V	Human, G, D	2,899	0.458	54.7	53.4	0.435	0.158
AL	*Ascaris lumbricoides*	III	Human, G, D	502	0.485	54.7	54.6	0.300	0.150
AS	*Ascaris suum*	III	Pig, G, D	5,813	0.431	55.1	56.1	0.336	0.109
BM	*Brugia malayi*	III	Human, G, OV	4,244	0.302	49.9	56.3	0.279	0.066
CE	*Caenorhabditis elegans*	V	Free living, A, n.a.	14,527	0.362	50.2	56.4	0.389	0.324
DI	*Dirofilaria immitis*	III	Canine, G, OV	1,152	0.277	48.9	56.8	0.246	0.053
GP	*Globodera pallida*	IVb	Potato, G, D	2,090	0.596	51.4	49.1	0.402	0.088
GR	*Globodera rostochiensis*	IVb	Potato, G, D	2,490	0.606	52.7	50.2	0.425	0.149
HC	*Haemonchus contortus*	V	Sheep/goat, G, D	4,003	0.407	55.5	56.5	0.378	0.113
HG	*Heterodera glycines*	IVb	Soya, G, D	7,427	0.587	51.8	50.7	0.415	0.094
HS	*Heterodera schachtii*	IVb	Beet, G, D	1,050	0.581	51.5	50.9	0.397	0.079
LS	*Litomosoides sigmodontis*	III	Rodent, G, OV	1,352	0.365	52.8	56.8	0.323	0.108
MA	*Meloidogyne arenaria*	IVb	Plants, OP, D	1,799	0.206	42.8	53.2	0.334	0.084
MC	*Meloidogyne chitwoodi*	IVb	Plants, FP, D	2,378	0.173	41.8	52.6	0.241	0.065
MH	*Meloidogyne hapla*	IVb	Plants, OP, D	4,699	0.200	42.1	52.6	0.211	0.082
MI	*Meloidogyne incognita*	IVb	Plants, OP, D	4,656	0.228	44.4	54.2	0.325	0.093
MJ	*Meloidogyne javanica*	IVb	Plants, OP, D	2,399	0.224	43.3	53.5	0.250	0.084
MP	*Meloidogyne paranaensis*	IVb	Plants, ?, D	1,080	0.217	43.2	53.1	0.275	0.114
NA	*Necator americanus*	V	Human, G, D	1,926	0.412	56.3	57.2	0.408	0.160
NB	*Nippostrongylus brasiliensis*	V	Rodent, G, D	639	0.505	54.2	51.5	0.448	0.274
OV	*Onchocerca volvulus*	III	Human, G, OV	2,811	0.321	50.2	56.6	0.276	0.071
OO	*Ostertagia ostertagi*	V	Cattle, G, D	1,732	0.427	55.3	55.6	0.382	0.101
PE	*Pratylenchus penetrans*	IVb	Plants, G, D	348	0.442	51.9	54.7	0.285	0.109
PV	*Pratylenchus vulnus*	IVb	Plants, G, D	526	0.589	50.9	50.1	0.340	0.101
PP	*Pristionchus pacificus*	V	Free living, G, n.a.	3,222	0.514	47.4	45.3	0.489	0.365
RS	*Radopholus similis*	IVb	Plants, G, D	305	0.635	49.6	45.4	0.358	0.116
SR	*Strongyloides ratti*	IVa	Rodent,[d] G/OP, D	2,923	0.099	35.8	46.3	0.380	0.257
SS	*Strongyloides stercoralis*	IVa	Human,[d] G/OP, D	2,910	0.124	37.3	47.4	0.391	0.239
TD	*Teladorsagia circumcincta*	V	Sheep, G, D	1,376	0.436	55.6	55.7	0.381	0.115
TC	*Toxocara canis*	III	Canine, G, D	1,048	0.458	55.0	54.9	0.326	0.131
TS	*Trichinella spiralis*	I	Mammals, G, D/P	2,772	0.364	50.7	56.8	0.263	0.044
TM	*Trichuris muris*	I	Mouse, G, D	1,085	0.518	54.6	52.6	0.323	0.087
TV	*Trichuris vulpis*	I	Canine, G, D	760	0.503	54.9	53.8	0.231	0.064
WB	*Wuchereria bancrofti*	III	Human, G, OV	1,176	0.333	51.0	56.1	0.362	0.112
XI	*Xiphinema index*	I	Plants, G, D	3,646	0.513	51.0	52.8	0.436	0.107
ZP	*Zeldia punctata*	IVb	Free living, G, n.a.	167	0.318	47.1	53.2	0.352	0.386

NOTE.—n.a., not available.

[a] From Blaxter et al. (1998) and Parkinson, Mitreva, et al. (2004).

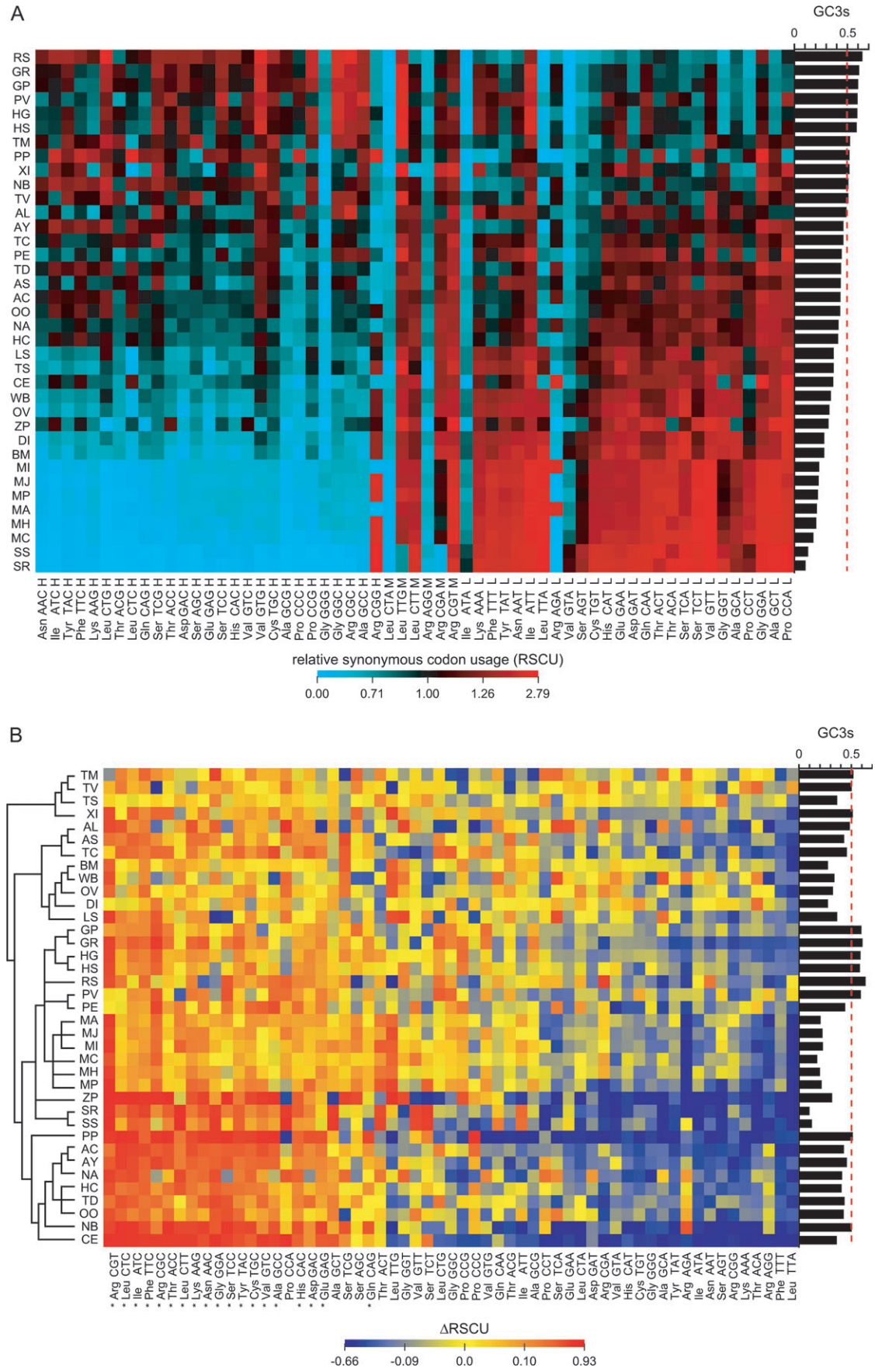
[b] G = gonochoric, A = androdioecious, OP = obligate parthenogen, FP = facultative parthenogen, D = direct transmission, D/P = transmission direct and via paratenic hosts, OV = obligate vector, ? = unknown.

[c] ≥100 codons long.

[d] Experiences a free-living stage.

of each codon relative to that expected under equal usage of alternative codons of the same amino acid. Heat maps of RSCU were constructed with CIMMiner (http://discover.nci.nih.gov/cimminer) (Weinstein et al. 1997). For several analyses, we partitioned loci by the observed counts of ESTs to define expression levels as low ($n = 1$), medium ($1 < n < n_{90}$), and high ($n \geq n_{90}$), where n_{90} is the species-specific 90th percentile count of ESTs (n_{90} ranged from 2 to 8; *C. elegans* $n_{90} = 38$).

Putative optimal codons were inferred for each species based on departures from equal codon usage by sets of loci with high and low gene expression (ΔRSCU), as inferred from EST counts (Duret and Mouchiroud 1999). ΔRSCU for a given codon is the difference between the average RSCU of genes with high and low expression (significance tested using 1-way analysis of variance (ANOVA) in JMP v5.0). We used the putatively optimal codons identified by this ΔRSCU analysis to compute F_{op}, using either the species-specific set of optimal codons or a consensus set of optimal codons (F_{cop}). In calculation of *C. elegans* F_{op}, we used the standard set of optimal codons previously described for this species (Stenico et al. 1994). We found that alternative approaches to identifying optimal codons, as implemented in CodonW (J Peden, http://codonw.sourceforge.net) and codbiasML (Slatkin and Novembre 2003; Wall and Herbeck 2003) did not satisfactorily separate the potential effects of selection from base composition, yielding sets of putatively optimal codons that closely mirrored the sets of codons with high overall RSCU in fig. 1 (i.e., major codons). In the case of correspondence analysis, this is due to the confounding effect of GC content on ENC because codonW uses ENC to partition genes rather than a more direct measure of gene expression. We follow the distinction of previous studies between major and optimal codons (Duret and Mouchiroud 1999; Kliman et al. 2003), where major codons exhibit RSCU > 1 and optimal codons have ΔRSCU > 0 at $P < 0.05$. Optimal codons were mapped onto the nematode phylogeny in Mesquite


A
GC3s
0
0.5
RS
GR
GP
PV
HG
HS
TM
PP
XI
NB
TV
AL
AY
TC
PE
TD
AS
AC
OO
NA
HC
LS
TS
CE
WB
OV
ZP
DI
BM
MI
MJ
MP
MA
MH
MC
SS
SR
Asn AAC H
Ile ATC H
Tyr TAC H
Phe TTC H
Lys AAG H
Leu CTG H
Thr ACG H
Leu CTC H
Gln CAG H
Ser TCG H
Thr ACC H
Asp GAC H
Ser AGC H
Glu GAG H
Ser TCC H
His CAC H
Val GTC H
Val GTG H
Cys TGC H
Ala GCG H
Pro CCC H
Pro CCG H
Gly GGG H
Gly GGC H
Arg CGC H
Ala GCC H
Arg CGG H
Leu CTA M
Leu TTG M
Leu CTT M
Arg AGG M
Arg CGA M
Arg CGT M
Ile ATA L
Lys AAA L
Phe TTT L
Tyr TAT L
Asn AAT L
Ile ATT L
Leu TTA L
Arg AGA L
Val GTA L
Ser AGT L
Cys TGT L
His CAT L
Glu GAA L
Asp GAT L
Gln CAA L
Thr ACT L
Thr ACA L
Ser TCA L
Ser TCT L
Val GTT L
Gly GGT L
Ala GCA L
Pro CCT L
Gly GGA L
Ala GCT L
Pro CCA L
relative synonymous codon usage (RSCU)
0.00
0.71
1.00
1.26
2.79
B
GC3s
0
0.5
TM
TV
TS
XI
AL
AS
TC
BM
WB
OV
DI
LS
GP
GR
HG
HS
RS
PV
PE
MA
MJ
MI
MC
MH
MP
ZP
SR
SS
PP
AC
AY
NA
HC
TD
OO
NB
CE
* Arg CGT
* Leu CTC
* Ile ATC
* Phe TTC
* Arg CGC
* Thr ACC
* Leu CTT
* Lys AAG
* Asn AAC
* Gly GGA
* Ser TCC
* Tyr TAC
* Cys TGC
* Val GTC
* Ala GCC
* Pro CCA
* His CAC
* Asp GAC
* Glu GAG
Ala GCT
Ser TCG
Ser AGC
* Gln CAG
Thr ACT
Leu TTG
Gly GGT
Val GTT
Ser TCT
Leu CTG
Gly GGC
Pro CCG
Pro CCC
Val GTG
Gln CAA
Thr ACG
Ile ATT
Ala GCG
Pro CCT
Ser TCA
Glu GAA
Leu CTA
Asp GAT
Arg CGA
Val GTA
His CAT
Cys TGT
Gly GGG
Ala GCA
Tyr TAT
Arg AGA
Ile ATA
Asn AAT
Ser AGT
Arg CGG
Lys AAA
Thr ACA
Arg AGG
Phe TTT
Leu TTA
ΔRSCU
-0.66
-0.09
0.0
0.10
0.93

v. 1.06 with ancestral states inferred by parsimony (http://mesquiteproject.org/mesquite/mesquite.html). We also created the new statistic $\overline{\Delta RSCU_+}$ to summarize codon bias for comparison among species, where $\overline{\Delta RSCU_+}$ is the average of all positive ΔRSCU values across codons within a species. Because RSCU is independent of amino acid content and ΔRSCU should control for base composition differences among genomes (Stenico et al. 1994; Duret and Mouchiroud 1999), $\overline{\Delta RSCU_+}$ is likely to be useful for comparing codon-bias information for different taxa that use different sets of genes.

We tested for evidence of an effect of natural selection in shaping codon-bias patterns by identifying significant Spearman rank correlation coefficients (ρ) between measures of codon bias and gene expression (as estimated from counts of ESTs) or base composition (third-position silent G + C content, GC3s) using the R statistical package (http://www.r-project.org). Because EST data do not provide noncoding DNA for most genes to allow inference of background base composition, we rely on GC3s as an index of base composition. GC3s was calculated with INCA from 4-fold silent sites (Supek and Vlahovicek 2004). To infer the relative importance of neutral and selective processes in shaping codon-usage bias of each species, we constructed ANOVA models in JMP v. 5 for codon-usage bias (F_{op}) as a function of base composition (GC3s), expression level ($\log_{10}$-transformed EST counts), EST length ($\log_{10}$ transformed), and all pairwise interactions.

Amino acid frequencies were calculated for each gene, along with the fraction of GC-rich and GC-poor amino acids defined previously as FYMINK (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine) and GARP (glycine, alanine, arginine, and proline), respectively (Foster et al. 1997). Amino acid frequencies were then used to test for differential effects of base composition and gene expression on protein-level characteristics using Spearman rank correlations and 1-way ANOVA.

Molecular Phylogeny of 37 Nematode Species

Based upon the data set from Blaxter et al. (1998), we estimated the phylogenetic relationships of the 37 species using an alignment of nuclear small subunit ribosomal RNA genes to place taxa absent from previous phylogenetic studies. The alignment was analyzed in PAUP v.4b.10 (Swofford 2001) using the Neighbor-Joining method and a General Time Reversible + G + I model of sequence evolution selected as best describing the data by Modeltest 3.0 (Posada and Crandall 1998). The robustness of the phylogeny was assessed by 1,000 bootstrap replicates, and nodes with support less than 70% collapsed to form polytomies. Where terminal nodes overlap, the phylogeny agrees with that defined previously (Blaxter et al. 1998) and confirmed in a more recent and comprehensive analysis (Meldal et al. 2006). The phylum can be divided into 5 major clades (termed clades I, II, III, IV, and V; clade II is not sampled here), which diverged approximately 700 MYA (Blaxter 1998). All members of clade III are parasitic, but the representatives of clades IV and V analyzed here include both free-living and parasitic species. Although many members of clade I are nonparasitic, only animal and plant parasites are included in this study. Based on this phylogeny, we used COMPARE to conduct phylogenetic mixed model (PMM) analyses of interspecific trait variation (Lynch 1991; E. Martins, http://compare.bio.indiana.edu). We generated 50 random topologies concordant with the polytomous nodes, using default parameters in COMPARE, to account for uncertainty in the tree; we report the resulting phylogenetic and ahistorical trait correlations.

Results

Base Composition and Gene Expression Both Affect Synonymous Codon Usage

An unrivalled resource of genomic data in the form of EST data sets is available for the phylum Nematoda, comprising a collection of 37 species that span its phylogenetic diversity (table 1; fig. 1). Our analysis incorporates an average of 2,284 genes per species (excluding *C. elegans*), each at least 100 amino acids long and with an average of 3.0 EST hits. Codon usage is highly nonrandom for all 37 nematode taxa (including *C. elegans*), and these species also differ dramatically in overall base composition, ranging from an average of 10–63% G + C bases at 4-fold silent sites (GC3s) (table 1; fig. 1). It is clear that base compositional differences among species contributes, at least in part, to their different relative usage of synonymous codons, with alternative codons with more G or C bases being incorporated relatively more frequently in high G + C content genomes (and vice versa for low G + C content genomes; fig. 1). However, we also find that many nematode species show significant codon-usage differences between genes from high and low classes of gene expression (fig. 2; similar results are observed for codon-bias indices other than N'_c). Likewise, codon bias (F_{op}) correlates positively with expression levels for many taxa independently of base composition, which is expected if selection for translational efficiency and accuracy contributes to codon bias (fig. 2).

Identification and Analysis of Optimal Codons

Given the inference that both neutral and selective forces shape codon-usage patterns, we identified putatively optimal codons. We calculated the RSCU for each codon in each gene of a given species and tested for a difference between those genes with high and low EST counts (ΔRSCU; Duret and Mouchiroud 1999); we considered as optimal

←

FIG. 1.—Heat map of (*A*) RSCU and (*B*) ΔRSCU values for 37 species of nematode. Each column represents a different codon, with the corresponding amino acid abbreviations and codon identity. Also indicated along the bottom: (*A*) the relative G + C content of synonymous alternative codons (H = high, M = moderate, and L = low) and (*B*) consensus optimal codons identified with an asterisk. Different species are represented in each row (identifiers as in table 1), sorted by (*A*) base composition (mean GC3s) or (*B*) by the phylogenetic topology indicated to the left. Significantly positive values of ΔRSCU are indicated by the optimal codons in figure 3.

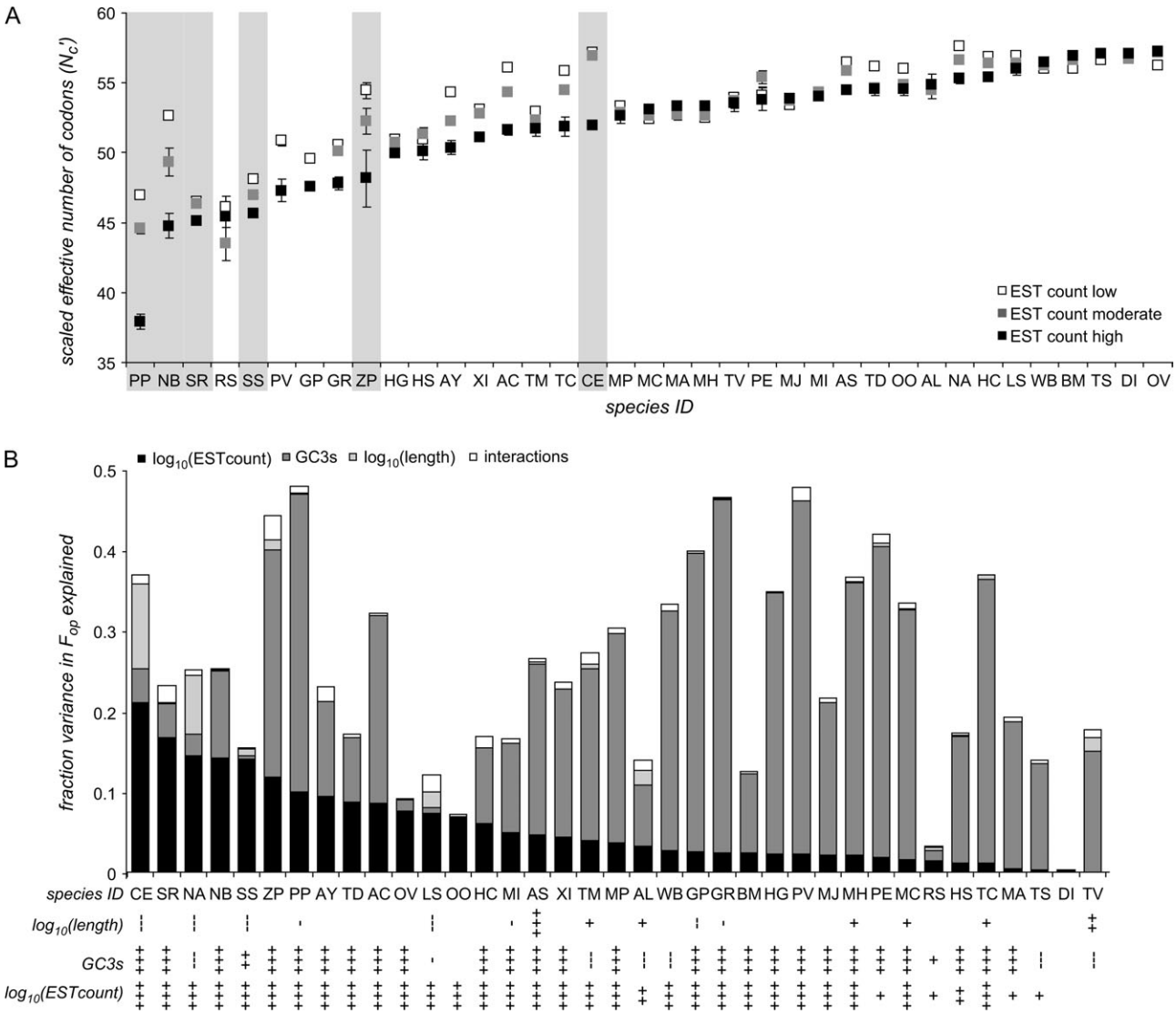


FIG. 2.—Association between codon-usage bias and gene expression. (*A*) Average N'_c for genes with low, medium, or high EST counts; the 6 species with high mean $\Delta RSCU_+$ are highlighted in gray. Error bars indicate ± 1 standard error. (*B*) The fraction of variance in the frequency of species-specific optimal codons (F_{op}) explained by different variables in multivariate analyses. Species are sorted by (*A*) increasing average N'_c for high EST-count genes and (*B*) decreasing influence of gene expression on F_{op}. Signs in (*B*) correspond to positive (+) or negative (−) associations, with the number of symbols indicating significance levels as +/− $P < 0.05$, ++/−− $P < 0.001$, +++/−−− $P < 0.0001$. Species identifiers as in table 1.

those codons with significantly higher RSCU among genes with high EST counts. The resulting putatively optimal codons for each nematode species are summarized in figure 3, and figure 1*B* gives a graphical representation of the continuous range of ΔRSCU values. Nineteen "consensus" optimal codons were observed across many species, including codons for all degenerate amino acids except proline, plus 2 codons for each of the 6-fold degenerate amino acids leucine and serine (fig. 3). These 19 consensus optimal codons overlap completely with the optimal codons described previously for *C. elegans*, lacking only the proline CCA, alanine GCT, and serine TCT codons (Stenico et al. 1994). For *C. elegans*, the ΔRSCU approach identifies the previously derived set of optimal codons (Stenico et al. 1994), plus the TCG codon of serine, to have significantly greater representation among highly expressed genes. To summarize consistency with the 19 consensus codons, we introduce 2 simple indices: p_c, the fraction of the consensus codons identified as optimal in a given species, and p_t, the fraction of the total number of optimal codons in a species that are consensus optimal codons. Those taxa showing the greatest consistency with the consensus optimal codons (high p_c) also have the most optimal codons identified ($\rho = 0.96$, $P < 0.0001$; PMM phylogenetic correlation = 0.12, ahistorical correlation = 0.94; supplementary fig 1, Supplementary Material online), suggesting that 1) the 19 consensus codons likely represent close to the full complement of optimal codons in these taxa, and 2) even deeply divergent nematodes have relatively similar sets of optimal codons.

The number of optimal codons identified in a species depends strongly on the number of genes represented in the

FIG. 3.—Optimal codons as identified by ΔRSCU analysis. Nineteen consensus optimal codons are indicated in gray. Species are sorted by the phylogenetic topology indicated to the left. * $P < 0.05$, ** $P < 0.001$, *** $P < 0.0001$, not significant.

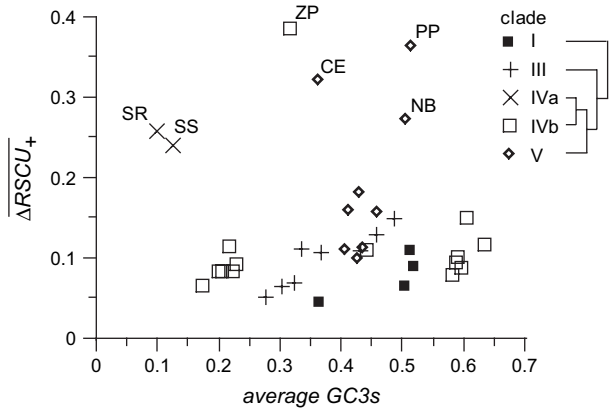


FIG. 4.—Differences among species in selection on codon usage. Average of positive ΔRSCU values per species indicate that 6 species have particularly strong selection on codon bias, spanning low, medium, and high GC-content genomes. Symbols indicate different clades within the nematode phylogeny.

sample (Spearman's $\rho = 0.70$, $P < 0.0001$; PMM phylogenetic correlation = 0.10, ahistorical correlation = 0.55), indicating that the power to detect putatively optimal codons is in part limited by sample size. However, p_t shows no strong association with gene number (PMM phylogenetic correlation = 0.04, ahistorical correlation = 0.17), with mean p_t highest in clade IV and clade V nematodes and lowest for species in clades I and III. Analyses using ANOVA with clade affiliation as a covariate give similar results (not shown). Thus, 1) the codons identified as optimal in taxa with few genes represented may not correspond to the full complement of optimal codons in those species and 2) the consensus optimal codons are primarily indicative of species in clades IV and V. Putative evolutionary changes in optimal codon identity are represented in the phylogenetic character mapping of optimal codons (supplementary fig. 2, Supplementary Material online), although the issue of sample size must also be considered when attempting to infer loss of optimal codons.

Differences in Codon-Usage Bias among Species

Given the identities of putatively optimal codons, we computed F_{op} and F_{cop}, the frequencies of species-specific optimal and consensus optimal codons, respectively (table 1; Ikemura 1985). Among the various codon-bias indices (including ENC and N'_c), F_{op} correlates least with GC3s (PMM phylogenetic correlation = −0.02, ahistorical correlation = 0.46; supplementary fig. 3, Supplementary Material online); consequently, we prefer F_{op} as a summary of selection on codon usage within a species. However, for comparing among taxa, averages of all of these codon-bias statistics give a poor indication of overall selection on codon usage for a species, due to covariation with base composition (supplementary fig. 3, Supplementary Material online). As an alternative, we consider average within-species ΔRSCU as an index of the strength of selection on codon usage for comparisons among taxa ($\overline{\Delta RSCU_+}$) and identify 6 outlier species with a particularly strong evidence of selection on codon usage (CE, PP, NB, SR, SS, and ZP; fig. 4).

In an effort to partition the variation in codon usage among loci into independent components associated with selective and nonselective factors, we constructed ANOVA models to describe intraspecific variation in F_{op} as a function of base composition (GC3s), gene expression (counts of ESTs), EST length, and their interactions. For 35 of the 37 species, codon-usage bias showed significant independent associations with gene expression in the direction predicted by the action of selection on codon usage (fig. 2*B*). However, base composition explains a much greater fraction of the variation in codon bias for many species than does gene expression (fig. 2*B*). Among those species with a strong effect of gene expression, EST length was frequently negatively associated with codon bias, whereas a positive correlation with length was more common amongst species with a weak correlation between codon-bias and expression level (fig. 2*B*). Pairwise interaction terms also contributed significantly to variation in codon-usage bias in some species, indicating that variation in the frequency of optimal codons is not always explained by a simple combination of factors. Although most of the species that show a large fraction of their variance in F_{op} explained by EST abundance in multivariate ANOVA tests also exhibit strong consistency with the 19 consensus codons (e.g., NB, PP, AY, NA, and AC), some species with only a weak effect of EST abundance on F_{op} also identify most of the same 19 consensus codons as optimal by the ΔRSCU analysis (e.g., HG, GR, and MH). Thus, correlations between F_{op} and gene expression do not necessarily capture a complete picture of the role of selection on codon usage. This is partly due to the ANOVA approach being unable to perfectly disentangle the issue of base composition because optimal codons tend to be GC rich and noncoding sequence is unavailable to accurately quantify local background GC content (Marais and Duret 2001); indeed, some studies have used GC3s itself as an index of codon bias (Tiffin and Hahn 2002; Wright et al. 2002). Consequently, selection may be the source of a portion of the variation in F_{op} that is explained by GC3s.

Nonrandom Amino Acid Usage

The relative abundance of amino acids that are rich in guanine and cytosine (glycine, alanine, arginine, and proline; GARP amino acids) is low within GC-poor nematode genomes, whereas such genomes show a high relative abundance of amino acids that are rich in adenine and thymine (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine; FYMINK amino acids) (GARP × GC3s PMM phylogenetic correlation = −0.11, ahistorical correlation = −0.79; FYMINK × GC3s PMM phylogenetic correlation = 0.20, ahistorical correlation = 0.79; fig. 5*A*). These associations also are evident within species (low-GC genes exhibit reduced GARP levels and elevated levels of FYMINK amino acids; PMM phylogenetic correlation = −0.13, ahistorical correlation = −0.86; fig. 5*B*). Thus, patterns of base composition within and between genomes influence patterns of amino acid usage, in addition to synonymous codon usage, among the species included in these analyses. The amino acid composition of genes also varies as a function of gene expression, such that some

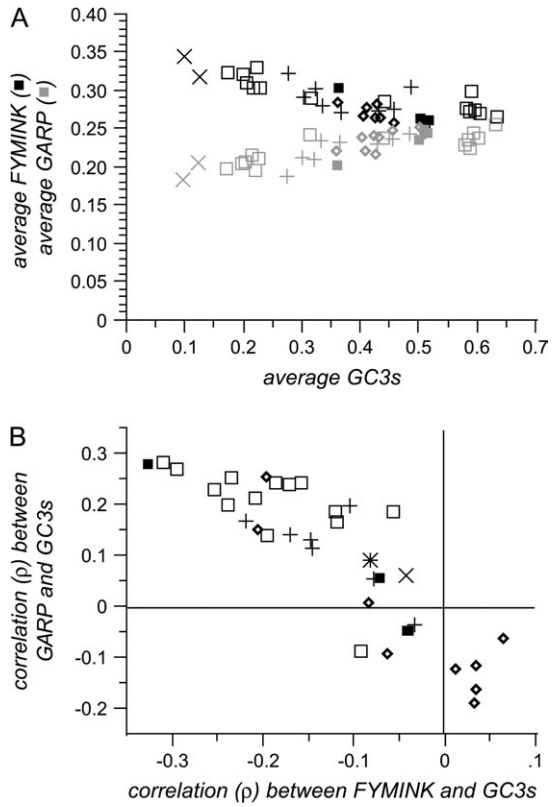


FIG. 5.—The influence of base composition on amino acid usage. (*A*) Average fraction FYMINK or GARP amino acids for each species. (*B*) Plot of the within-species correlation coefficients (Spearman's ρ) between GC3s and the fraction of either FYMINK or GARP amino acids. Symbols indicate different clades within the nematode phylogeny as in figure 4.

amino acids tend to be more abundant (e.g., Gly, Ala, and Lys) or less abundant (e.g., Ser, Leu, Phe, Ile, and Asn) in genes with many ESTs (supplementary fig. 4, Supplementary Material online). This can also be quantified in terms of the average $\Delta RSCU_+$ per amino acid for each species, which indicates that some amino acids (mainly the highly degenerate amino acids) tend to exhibit more strongly biased codon-usage patterns in highly expressed genes than do other amino acids (e.g., Arg, Leu, and Ser; supplementary fig. 5, Supplementary Material online). However, it is unclear whether these observations reflect different selective costs of functionally similar amino acids, variation in the abundance of protein classes with different peptide domain characteristics among highly and lowly expressed genes, or a combination of factors.

Discussion

Neutral and Selective Forces Shape Codon Usage in Nematodes

Selection for translational efficiency and/or accuracy has long been believed to be a cause of codon-usage biases in the *C. elegans* genome (Stenico et al. 1994), with supporting evidence from diverse data sets (Duret and Mouchiroud 1999; Duret 2000; Marais and Duret 2001; Castillo-Davis and Hartl 2002; Cutter et al. 2003; Cutter and Ward 2005). Here we show that such selection on codon bias extends to diverse members of the phylum Nematoda. In addition, the local base composition of genes and the overall pattern of base composition in a genome contribute to variation in codon-usage bias within and between nematode species: the stronger the skew in base composition, the greater the bias in codon usage. We also demonstrate that previous observations of stronger codon bias in short genes (Moriyama and Powell 1998; Duret and Mouchiroud 1999; Coghlan and Wolfe 2000) is repeated in several species of nematodes, particularly among those that have a strong influence of gene expression on their patterns of codon usage. However, we emphasize that it is not appropriate to infer the relative strength of selection among species using average ENC or F_{op} because of their covariation with base composition or use of different sets of optimal codons (Comeron and Aguade 1998; Herbeck and Novembre 2003)(supplementary fig. 3, Supplementary Material online). We propose to quantify the importance of selection on codon usage among species using the relative values of ΔRSCU averaged across amino acids, although this $\overline{\Delta RSCU_+}$ statistic also may be an imperfect index. Most nematodes with evidence for adaptive codon bias preferentially utilize a consensus set of codons in genes with high expression, although phylogenetic history and skewed genomic base composition appear to play a role in the evolution of some alternative optimal codons. Among these 37 species, exhibiting a very wide range of average GC content, it is important to differentiate between codons that are used more often overall (major codons) from those that differ in abundance in relation to gene expression (optimal codons) because major codons are strongly influenced by base composition and frequently are not identified as optimal.

Alternative Sets of Optimal Codons

The collection of inferred optimal codons for most species corresponds to a set of 19 consensus optimal codons for 17 amino acids. In the case of 5 amino acids, none of the 37 species exhibits a preference for the alternative codon (fig. 3, supplementary fig. 2, Supplementary Material online). This trend illustrates the impressive consistency in optimal codon identities across hundreds of millions of years of nematode evolution, as has also been suggested in bacteria, yeast, and *Drosophila* (Ikemura 1985; Kreitman and Antezana 1999). However, the sets of optimal codons for all species deviate from the consensus in one or more ways: 1) the identity of the optimal codon has switched to an alternative degenerate codon, 2) an additional optimal codon increases the number of optimal codons for an amino acid, and 3) no optimal codon is present for a given amino acid. In those species with strong evidence of selection on codon usage, it is reasonable to ascribe differences from the consensus optimal codon set to evolutionary processes (e.g., gain of proline CCC and serine TCT in *Pristionchus pacificus*, switch to alanine GCG and serine TCG in *Heterodera glycines*). In particular, such shifts may indicate selection-shaping changes in codon preference in association with differences in effective population size (Kreitman and Antezana 1999). We also speculate that the extreme base composition bias toward A/T in the 2 *Strongyloides* species might have contributed a selective force involved

in switches in optimal codons for glutamic acid (CAG to CAA) and proline (CCC to CCA). Studies of single organelle genes in large collections of insect and plant taxa similarly found relatively few transitions in optimal codon identity, with shifts involving 2 preferred codons in 4- and 6-fold degenerate amino acids being more prevalent than shifts between alternative 2-fold degenerate codons (Herbeck and Novembre 2003; Wall and Herbeck 2003).

Putatively optimal codons also are missing for many amino acids in some species. For some cases, this probably reflects limited power to identify optimal codons due to small sample size of genes sequenced (e.g., HS and PV), whereas for other species for which many genes were included in analysis, selection may be unable to distinguish between alternative codons in some amino acids with particularly weak selection (e.g., TS, MC, BM, and OV). Small effective population size might allow genetic drift to lead to shifts in codon preference and, more generally, eliminate patterns of codon preference (Kreitman and Antezana 1999). Differences in the isoaccepting tRNA pools within cells during different stages of development also could weaken selection for codon bias (Moriyama and Powell 1997). We infer that there is no role of selection-shaping patterns of codon bias in species with only a few putatively optimal codons that differ from the consensus set with low statistical support (e.g., TV, TS, DI, RS, and WB). Additionally, species with few genes analyzed must await further data for a final determination of the full complement of optimal codons (e.g., ZP).

Several codons were universally underrepresented across species (arginine AGG, glycine GGG, isoleucine ATA, leucine CTA, and valine GTA). The glycine GGG codon is also rarely used in *Drosophila* species and *Escherichia coli*, probably due to a detrimental effect on mRNA tertiary structure (Kreitman and Antezana 1999). However, it is less clear why the other codons are so rare in both absolute terms and especially in highly expressed genes.

Differences in codon usage for several amino acids reflect an effect of phylogeny. For example, all *Meloidogyne* species and most Spiruromorph nematodes (including *Brugia malayi*) use the leucine TTG as an optimal codon, whereas their nearest outgroup species do not. By contrast, ahistorical features also contribute to alternative codon preferences. For example, several unrelated low-GC genomes preferentially use isoleucine ATT and threonine ACT codons, unlike their nearest relatives with higher GC-content. Optimal codon changes among species for alanine and threonine illustrate the potential for both phylogeny and base composition to affect the loss, gain, and switching of optimal codon identities (fig. 6, supplementary fig. 2, Supplementary Material online), although the long phylogenetic timescale and predominance of parasitic species in this data set makes any inference of ancestral states preliminary.

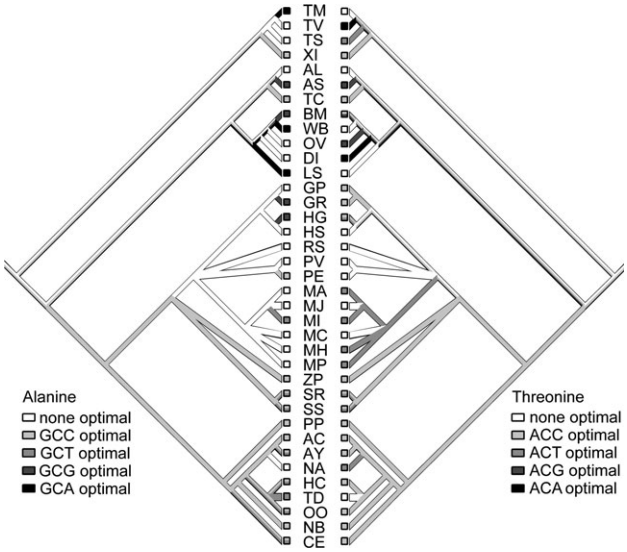


FIG. 6.—Mapping of optimal codons for alanine and threonine on the nematode phylogeny with ancestral states inferred by parsimony. See supplementary figure 2, Supplementary Material online for character maps of all amino acids.

Nonrandom Patterns of Amino Acid Usage

In addition to affecting codon-usage patterns, genomic base composition also influences amino acid usage in these nematode species. Specifically, the incidence of GC-poor amino acids is greater among proteins of species with overall low GC content (and vice versa for GC-rich amino acids; FYMINK × GC3s PMM phylogenetic correlation = 0.20, ahistorical correlation = −0.79; GARP × GC3s PMM phylogenetic correlation = −0.11, ahistorical correlation = 0.79). These findings are entirely consistent with previous reports for bacteria (Sueoka 1961; Gu et al. 1998; Singer and Hickey 2000), plants (Wang et al. 2004), and animals (D'Onofrio et al. 1991; Porter 1995; Foster et al. 1997). The problems that this may cause for phylogenetic reconstruction based on peptide alignments has long been noted (Steel et al. 1993), making appropriate models of nucleotide change an important feature of analyses of divergence and gene prediction.

We also report that certain amino acids are more common among highly expressed genes, as has been shown previously in bacteria (Akashi and Gojobori 2002; Merkl 2003). It is tempting to apply an adaptationist explanation to this pattern, such that overrepresented amino acids might be metabolically less costly (Akashi and Gojobori 2002) or have correspondingly higher tRNA abundances, permitting greater translational efficiency or accuracy. However, it will be important to rule out the possibility that this pattern simply reflects base composition effects or the kinds of genes that are expressed at high levels (e.g., multigene families and classes of genes with similar domain structures) before concluding that some amino acids confer a selective advantage when incorporated into abundant proteins in place of functionally equivalent amino acids. Nevertheless, the propensity for optimal codons to be identified more frequently for some amino acids (e.g., Phe vs. Gln, Thr vs. Pro, and Leu vs. Ser) and for the magnitude of ΔRSCU to be greater for some amino acids than others (e.g., Arg, Leu, and Ser) suggests that the strength of selection does differ among amino acids, perhaps reflecting a "hierarchy of selection coefficients" (McVean and Vieira 2001). Similar variation among amino acids in *E. coli* and in *Drosophila* species has been interpreted as evidence of different strengths of

selection for optimal codons in different amino acids (Moriyama and Powell 1997; McVean and Vieira 2001; Fuglsang 2003).

Selection on Codon Usage: Life History Characters and Population Genetic Implications

Life history characteristics are known to contribute to differences in codon-usage patterns in bacteria and archaea. For instance, thermophilic and mesophilic species exhibit different patterns independently of base compositional effects (McDonald 2001; Carbone et al. 2005). However, comparable discrepancies associated with life history have been less forthcoming in eukaryotes, for example, in terms of the expected differences for species with alternative modes of reproduction (Tiffin and Hahn 2002; Wright et al. 2002). The nematode species considered in this study differ in life history along several axes, including parasitism, host specificity, and mode of reproduction. We observe no obvious pattern associated with host specificity or breeding system, in contrast to the incidence of a parasitic versus free-living lifestyle. Only 3 species in this data set are free living (PP, ZP, *C. elegans*), and all 3 demonstrate robust evidence for selection on codon-usage bias, compared with only 3 of 35 parasitic species (fig. 4). Furthermore, of these 3 parasitic species, the 2 *Strongyloides* species are unusual in that they have a free-living stage (Viney 1999). Species with larger effective population sizes are expected to exhibit stronger adaptive bias among codons. This suggests that nematodes with obligate or facultative free-living life histories may in general have larger effective population sizes than obligate parasites and, additionally, that many obligate parasitic nematodes will not respond efficiently to the weak selection that acts on codon usage. *Nippostrongylus brasiliensis* also exhibits strong selection on codon-usage bias, yet this rat parasite does not have obvious features of lifestyle or abundance in the wild that that are known to differ from its close relatives (including the human hookworms and sheep barber pole nematode) that could explain this finding. However, it is important to point out that the selection differential between alternative codons in highly expressed genes is sufficient to allow detection of some optimal codons in most taxa, including parasites.

Given that natural selection contributes to nonrandom codon usage in nematodes, these data also inform questions relating to the relative strength of selection for efficient translation of different amino acids. McVean and Vieira (2001) incorporate the notion of a hierarchy of selection coefficients among amino acids into their models of selection on codon-usage bias. A hierarchy of selection coefficients would suggest that ΔRSCU will be greater for codons subject to stronger selection, so the ranking of codons in fig. 1*B* may provide a gauge of the relative strength of selection on different codons. To more completely dissect the role of selection in shaping codon-usage patterns, it would be ideal to obtain polymorphism data to quantify the strength of selection, as has been done for species of Drosophila (e.g., Hartle et al. 1994; Akashi 1995; McVean and Vieira 2001; Maside, Lee, and Charlesworth 2004), humans (Williamson et al. 2005), and the nematode C. remanei (Cutter and Charlesworth 2006).

Supplementary Material

Supplementary figures 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). No GenBank accession numbers are included.

Acknowledgments

We thank the Charlesworths' lab groups for constructive discussion of this work, A. Betancourt, D. Charlesworth, K. Wolfe and 3 reviewers for comments on the manuscript, and R. Schmid for access to and maintenance of NEMBASE. We also thank D. Gaffney for assistance with R. A.D.C. is supported by International Research Fellowship Program grant #0401897 from the National Science Foundation. J.D.W. is supported by the BBSRC.

Literature Cited

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA. Genetics. 139:1067–1076.

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA. 99:3695–3700.

Barrai I, Volinia S, Scapoli C. 1995. The usage of oligopeptides in proteins correlates negatively with molecular-weight. Int J Peptide Protein Res. 45:326–331.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. J Biol Chem. 257:3026–3031.

Blaxter ML. 1998. *Caenorhabditis elegans* is a nematode. Science. 282:2041–2046.

Blaxter ML, De Ley P, Garey JR, et al. (12 co-authors). 1998. A molecular evolutionary framework for the phylum Nematoda. Nature. 392:71–75.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics. 129:897–907.

Carbone A, Kepes F, Zinovyev A. 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol Biol Evol. 22:547–561.

Castillo-Davis CI, Hartl DL. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. Mol Biol Evol. 19:728–735.

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA. 101:3480–3485.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast. 16:1131–1145.

Comeron JM, Aguade M. 1998. An evaluation of measures of synonymous codon usage bias. J Mol Evol. 47:268–274.

Cutter AD, Charlesworth B. 2006. Selection intensity on preferred codons correlates with overall codon usage bias in Caenorhabditis remanei. Current Biology In press.

Cutter AD, Payseur BA, Salcedo T, et al. (12 co-authors). 2003. Molecular correlates of genes exhibiting RNAi phenotypes in *Caenorhabditis elegans*. Genome Res. 13:2651–2657.

Cutter AD, Ward S. 2005. Sexual and temporal dynamics of molecular evolution in *C. elegans* development. Mol Biol Evol. 22:178–188.

D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol. 32:504–510.

Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16:287–289.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.
Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proc Natl Acad Sci USA. 96:4482–4487.
Ewing B, Green P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. 8:186–194.
Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol. 44:282–288.
Fuglsang A. 2003. The effective number of codons for individual amino acids: some codons are more optimal than others. Gene. 320:185–190.
Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8:R49–R62.
Gu X, Hewett-Emmett D, Li WH. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica. 103:383–391.
Hartl DL, Moriyama EN, Sawyer SA. 1994. Selection intensity for codon bias. Genetics. 138:227–234.
Haywood-Farmer E, Otto SP. 2003. The evolution of genomic base composition in bacteria. Evolution. 57:1783–1792.
Herbeck JT, Novembre J. 2003. Codon usage patterns in cytochrome oxidase I across multiple insect orders. J Mol Evol. 56:691–701.
Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J Mol Biol. 158:573–597.
Ikemura T. 1985. Codon usage and transfer-RNA content in unicellular and multicellular organisms. Mol Biol Evol. 2: 13–34.
Kliman RM, Irving N, Santiago M. 2003. Selection conflicts, gene expression, and codon usage trends in yeast. J Mol Evol. 57:98–109.
Knight RD, Freeland SJ, Landweber LF. 2001. 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. 2:10.11–10.13.
Kreitman M, Antezana M. 1999. The population and evolutionary genetics of codon bias. In: Singh RS, Krimbas CB, editors. Evolutionary genetics: from molecules to morphology. New York: Cambridge University Press. p. 82–101.
Li WH. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol. 24:337–345.
Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene. 205:309–316.
Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution. 45:1065–1080.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330–338.
Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J Mol Evol. 52:275–280.
Maside XL, Lee AWS, Charlesworth B. 2004. Selection on codon usage in *Drosophila americana*. Curr Biol. 14:150–154.
McDonald JH. 2001. Patterns of temperature adaptation in proteins from the bacteria *Deinococcus radiodurans* and *Thermus thermophilus*. Mol Biol Evol. 18:741–749.
McVean GAT, Vieira J. 1999. The evolution of codon preferences in Drosophila: a maximum-likelihood approach to parameter estimation and hypothesis testing. J Mol Evol. 49: 63–75.
McVean GAT, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics. 157:245–257.
Meldal BHM, Debenham NJ, de Ley P, et al. (14 co-authors). Forthcoming. An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. Mol Biol Evol.
Merkl R. 2003. A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. J Mol Evol. 57:453–466.
Mitreva M, Blaxter ML, Bird DM, McCarter JP. 2005. Comparative genomics of nematodes. Trends Genet. 21:573–581.
Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol. 45:514–523.
Moriyama EN, Powell JR. 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. 26:3188–3193.
Morton BR, Levin JA. 1997. The atypical codon usage of the plant psbA gene may be the remnant of an ancestral bias. Proc Natl Acad Sci USA. 94:11434–11438.
Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 19:1390–1394.
Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M. 2004. PartiGene—constructing partial genomes. Bioinformatics. 20:1398–1404.
Parkinson J, Guiliano D, Blaxter M. 2002. Making sense of EST sequences by CLOBBing them. BMC Bioinformatics. 3:31.
Parkinson J, Mitreva M, Whitton C, et al. (12 co-authors). 2004. A transcriptomic analysis of the phylum Nematoda. Nat Genet. 36:1259–1267.
Parkinson J, Whitton C, Schmid R, Thomson M, Blaxter M. 2004. NEMBASE: a resource for parasitic nematode ESTs. Nucleic Acids Res. 32:D427–D430.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 23:301–309.
Porter TD. 1995. Correlation between codon usage, regional genomic nucleotide composition, and amino acid composition in the cytochrome P-450 gene superfamily. Biochim Biophys Acta Gene Struct Expr. 1261:394–400.
Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.
Seligmann H. 2003. Cost-minimization of amino acid usage. J Mol Evol. 56:151–161.
Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. 1995. DNA-sequence evolution—the sounds of silence. Philos Trans R Soc Lond Ser B Biol Sci. 349:241–247.
Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol. 4:222–230.
Singer GAC, Hickey DA. 2000. Nucleotide bias causes a genome-wide bias in the amino acid composition of proteins. Mol Biol Evol. 17:1581–1588.
Slatkin M, Novembre J. 2003. Appendix to paper by Wall and Herbeck—evolutionary patterns of codon usage in the chloroplast gene rbcL. J Mol Evol. 56:689–690.
Steel MA, Lockhart PJ, Penny D. 1993. Confidence in evolutionary trees from biological sequence data. Nature. 364: 440–442.
Stenico M, Lloyd AT, Sharp PM. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. 22:2437–2446.
Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA. 85:2653–2657.

Sueoka N. 1961. Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harbor Symp Quant Biol. 26:35–43.

Supek F, Vlahovicek K. 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics. 20:2329–2330.

Swofford D. 2001. PAUP 4b10 phylogenetic analysis using parsimony * and other methods. Sunderland (MA): Sinauer Associates.

Tiffin P, Hahn MW. 2002. Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp *pekinensis*. J Mol Evol. 54:746–753.

Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics. 159:1191–1199.

Viney ME. 1999. Exploiting the life cycle of *Strongyloides ratti*. Parasitol Today. 15:231–235.

Wall DP, Herbeck JT. 2003. Evolutionary patterns of codon usage in the chloroplast gene rbcL. J Mol Evol. 56:673–688.

Wang HC, Singer GAC, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. Mol Biol Evol. 21: 90–96.

Wasmuth J, Blaxter M. 2004. prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics. 5:187.

Weinstein JN, Myers TG, O'Connor PM, et al. (21 co-authors). 1997. An information-intensive approach to the molecular pharmacology of cancer. Science. 275:343–349.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA. 102:7882–7887.

Wright F. 1990. The effective number of codons used in a gene. Gene. 87:23–29.

Wright SI, Lauga B, Charlesworth D. 2002. Rates and patterns of molecular evolution in inbred and outbred Arabidopsis. Mol Biol Evol. 19:1407–1420.

Wright SI, Yau CBK, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol Biol Evol. 21:1719–1726.

Kenneth Wolfe, Associate Editor

Accepted August 23, 2006