Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes

Hong Qin*,[1], Wei Biao Wu†, Josep M. Comeron‡, Martin Kreitman*, and Wen-Hsiung Li*,[2]

**Department of Ecology and Evolution, †Department of Statistics, University of Chicago, Chicago, Illinois 60637, U.S.A. ‡Department of Biological Sciences, University of Iowa, Iowa City, IA 52242, U.S.A.*

Running Title: Intragenic spatial codon usage bias

Key words and phrases: spatial pattern, codon usage bias, Hill-Robertson effect, translational selection, GC content variation.

[1]Current address: Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642

[2]Corresponding author:
Wen-Hsiung Li
Department of Ecology and Evolution
University of Chicago
1101 East 57th Street
Chicago, IL 60637
Phone: 773-702-3104, Fax: 773-702-9740
Email: whli@uchicago.edu

ABSTRACT

To study the roles of translational accuracy, translational efficiency, and the Hill-Robertson effect in codon usage bias, we studied the intragenic spatial distribution of synonymous codon usage bias in four prokaryotic (*Escherichia coli, Bacillus subtilis, Sulfolobus tokodaii, and Thermotoga maritima*) and two eukaryotic (*Saccharomyces cerevisiae* and *Drosophila melanogaster*) genomes. We generated super-sequences at each codon position across genes in a genome and computed the overall bias at each codon position. By quantitatively evaluating the trend of the spatial pattern using isotonic regression, we show that in yeast and prokaryotic genomes, codon usage bias increases along translational direction. We argue that purifying selection against nonsense errors accounts for this incremental pattern. Fruit fly genes show a nearly symmetric M-shaped spatial pattern of codon usage bias, with less bias in the middle and both ends. The M-shaped pattern in fruit fly is stronger in intronless genes than in those with introns. The low codon usage bias in the middle region of fruit fly genes is best explained by interference (the Hill-Robertson effect) between selections at different codon positions. In both yeast and fruit fly, spatial patterns are characteristically different from patterns of GC-content variations. Effect of expression level on the strength of codon usage bias is more conspicuous than its effect on the shape of the spatial distribution.

CODON usage bias refers to the non-random usage of synonymous codons. Frequently used codons are often termed optimal or major codons, whereas less frequently used ones are termed nonoptimal or minor codons. Nonoptimal codons usually correspond to less abundant tRNAs than optimal codons (IKEMURA 1981; IKEMURA 1982; IKEMURA 1985; BULMER 1987; AKASHI 2001) and the translational machinery is more likely to stall there (KURLAND 1992). These processing errors, termed premature termination or nonsense errors, can occur at comparable levels to mis-incorporation, or mis-sense errors, during the elongation stage of translation (KURLAND 1992). Therefore, purifying selection is expected to increase in intensity as peptide elongation proceeds (EYRE-WALKER 1996b; AKASHI 2001) and should lead to gradual increase in codon usage bias along a gene (EYRE-WALKER 1996b). In addition, optimal codons may be selected for translational efficiency and so are more frequently used in highly expressed genes (SHARP and LI 1986; SHIELDS and SHARP 1987; SHARP *et al.* 1993; HARTL *et al.* 1994; MORIYAMA and POWELL 1997; AKASHI 2001; CARLINI and STEPHAN 2003). Therefore, codon usage bias may be accounted for by negative selection against nonoptimal codons and positive selection for optimal codons. Neither selection against mis-sense errors nor selection for translational efficiency is expected to be position dependent within genes, whereas selection against nonsense errors is expected to leave a incremental pattern of codon bias along translational direction within genes (EYRE-WALKER 1996b).

Recombination is another important but often controversial factor that affects codon usage bias. Selection at one locus can interfere with selection at another locus, known as the Hill-Robertson effect (HILL and ROBERTSON 1966). Negative selection against nonoptimal codons and positive selection for optimal codons are both likely to be weak (HARTL *et al.* 1994; AKASHI 1995; AKASHI and SCHAEFFER 1997; KLIMAN 1999; MASIDE *et al.* 2004). Nevertheless, selection

at one codon position within a gene can interfere with selection at other codon positions in the gene if they are tightly linked (LI 1987; MCVEAN and CHARLESWORTH 2000; COMERON and KREITMAN 2002). As a result, recombination is expected to increase the selection efficacy on codon usage bias. Several studies have observed a positive correlation between recombination rate and codon usage bias (KLIMAN and HEY 1993; COMERON *et al.* 1999; BETANCOURT and PRESGRAVES 2002; COMERON and KREITMAN 2002; HEY and KLIMAN 2002; KLIMAN and HEY 2003). However, there have also been observations suggesting that the Hill-Robertson effect has a minor influence on codon usage bias and is overwhelmed by the mutation pressure associated with recombination (MARAIS *et al.* 2001; MARAIS *et al.* 2003). The controversy around the Hill-Robertson effect on codon bias is partially due to the uncertainty in the recombination rate estimation (KLIMAN and HEY 2003; MARAIS *et al.* 2003). This uncertainty may be mitigated by studying the intragenic pattern of codon usage bias because intragenic variation of recombination rate is unlikely to be as extensive as intergenic variation. Comeron and Kreitman proposed that the Hill-Robertson effect should yield a distinctive intragenic spatial pattern of codon usage bias, with low bias in the middle but high bias near the ends (COMERON and KREITMAN 2002). In fruit fly genome, these authors showed that codon usage bias was indeed lower in the middle segment of long exons, and centrally located introns were found to mitigate the interference effect.

In short, the intragenic spatial distribution pattern of synonymous codon usage bias (spatial codon usage bias, for simplicity) may bear the signatures of selection against nonsense errors and the Hill-Robertson effect. This spatial bias pattern and other spatial patterns (such as substitution rate and GC-content variation) are important in molecular evolution and evolutionary genomics (BULMER 1991; MCVEAN and CHARLESWORTH 2000; COMERON and KREITMAN 2002; MARAIS and CHARLESWORTH 2003).

Spatial codon usage bias was first addressed in *E. coli* genes (BULMER 1988; CHEN and INOUYE 1990; EYRE-WALKER and BULMER 1993; CHEN and INOUYE 1994; EYRE-WALKER 1996a). Due to the nonlinear nature of the spatial patterns, most studies detected statistically significant changes only in regions near the start codons. Many studies also focused on the preferential usage of nonoptimal codons in the initial regions (BURNS and BEACHAM 1985; CHEN and INOUYE 1990; OHNO *et al.* 2001). A more sophisticated statistical method, such as isotonic regression, that can detect the trend of codon usage bias in the full length should be used. Isotonic regression refers to the test of an alternative hypothesis with ordered expectations (ROBERTSON *et al.* 1988; SOKAL and ROHLF 1995; WU *et al.* 2001). It is a very powerful method to test the existence of a monotonic trend in nonlinear data structure. For example, it has been successfully used in a global warming study (WU *et al.* 2001).

In eukaryotic genomes, spatial codon usage bias has received less attention (BULMER 1988; KLIMAN and EYRE-WALKER 1998; COMERON and KREITMAN 2002). Consequently, it is unclear as to the relative importance of selection against nonsense errors, selection against mis-sense errors, and selection for translational efficiency. It is even controversial whether linked weak selection has a discernable effect on codon bias (KLIMAN and HEY 1993; COMERON *et al.* 1999; MCVEAN and CHARLESWORTH 2000; MARAIS *et al.* 2001; COMERON and KREITMAN 2002; HEY and KLIMAN 2002; MARAIS and PIGANEAU 2002; KLIMAN and HEY 2003; KLIMAN *et al.* 2003; MARAIS *et al.* 2003). To address these issues, we present a systematic study of the intragenic spatial codon usage bias in both prokaryotic and eukaryotic genomes.

Codon usage bias may also be influenced by the interaction of mutation, selection, and random drift (BULMER 1987; SHIELDS and SHARP 1987; BULMER 1991; SHARP *et al.* 1993; KLIMAN and HEY 1994; EYRE-WALKER and BULMER 1995; AKASHI 1997; AKASHI *et al.* 1998;

RAND and KANN 1998), effective population size (LI 1987; BULMER 1991; BERG 1996), evolutionary history (BEGUN 2001), biased gene conversion (GALTIER *et al.* 2001; BIRDSELL 2002; GALTIER 2003; MARAIS 2003), mRNA secondary structure (EYRE-WALKER and BULMER 1993; HARTL *et al.* 1994; ANTEZANA and KREITMAN 1999; CHEN *et al.* 1999; CARLINI *et al.* 2001), translational initiation (SAKAI *et al.* 2001; STENSTROM *et al.* 2001; NIIMURA *et al.* 2003), and many other factors (FITCH 1980; MODIANO *et al.* 1981; BULMER 1990; BERG 1996; MCVEAN and HURST 2000; PLOTKIN and DUSHOFF 2003). These confounding factors make codon usage bias a difficult subject to study. Here, we focus on the intragenic spatial codon usage bias and study the relative strength of codon usage bias along the open reading frame. This approach enables us to detect the expected signatures of selection against nonsense errors and the Hill-Robertson effect.

MATERIALS AND METHODS

Genomic data

We parsed out 5366 reliable ORFs from a version of the *Saccharomyces cerevisiae* genome downloaded from the SGD website on July 11, 2003. The annotations of these 5366 ORFs have been confirmed by a comparative genome analysis (KELLIS *et al.* 2003). We excluded ORFs that were either questionable or with uncertain annotations.

The release 3.1 version of the *Drosophila melanogaster* genome was downloaded from Flybase on March 20, 2003. We parsed out 10617 ORFs with a single splicing isoform and restricted our analysis in these genes because genes with multiple splicing isoforms are more likely to be mis-annotated. From these single splicing isoforms, we parsed out 2171 intronless genes and 2049 genes with two coding exons.

The Genbank accession numbers for the prokaryotic genomes are: *Escherichia coli K12* (U00096), *Bacillus subtilis* (AL009126), *Sulfolobus tokodaii* (NC_003106), and *Thermotoga maritima* (AE000512).

In all genomes studied, we restricted our analysis to ORFs longer than 150 codons because short genes do not have enough codons for their overall pattern to be evaluated.

Gene expression data

Gene expression levels in *S. cerevisiae* were generated by the Young lab at MIT (http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt) (HOLSTEGE *et al.* 1998).

Gene expression levels in *D. melanogaster* were generated by the Reichert lab at the University of Basel. The Omnibus accession numbers are GPL70, GSM1359, GSM1360, GSM1361, and GSM1362 (MONTALTA-HE *et al.* 2002).

Isotonic regression

Isotonic regression is a method for fitting data to order constraints (ROBERTSON *et al.* 1988; WU *et al.* 2001). “Isotonic” means "order preserving". This non-parametric method is a very powerful way to test the existence of a monotonic trend under a nonlinear model. Suppose that we observe a sequence of random values, X_i, under the model $X_i = \mu_i + \sigma\varepsilon_i$, $1 \leq i \leq n$, where ε_i are independent standard normal random variables and $\sigma > 0$ is the standard deviation. We want to test the null hypothesis H_0: $\mu_1 = \mu_2 \ldots = \mu_n$ (constant) versus the alternative hypothesis H_A: $\mu_1 \leq \mu_2 \ldots \leq \mu_n$ with at least one i such that $\mu_i < \mu_{i+1}$ (a monotonic increasing trend). To define the isotonic test, assume at the outset that σ is known. The following likelihood-based test is proposed by (WU *et al.* 2001):

$$\Lambda_{n,r} = \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(\hat{\mu}_{k,r} - \overline{X}\right)^2 , \quad (1)$$

where $\overline{X} = \sum_{i=1}^{n} X_i / n$, r is a penalty factor that will be defined later, and $\hat{\mu}_{k,r}$ is the estimated trend

$$\hat{\mu}_{k,r} = \max_{i \le k} \min_{j \ge k} \frac{X_{i,r} + \ldots + X_{j,r}}{j - i + 1} \quad (2)$$

Here $X_{1,r} = X_1 + r\sqrt{n}$, $X_{n,r} = X_n - r\sqrt{n}$ and $X_{i,r} = X_i$ for $2 \le i \le$ n-1. The estimated trend (Eq. 2) is non-decreasing in *k*. Under the requirement $r > 0$, $\hat{\mu}_{1,r}$ and $\hat{\mu}_{n,r}$ are consistent estimators of μ_1 and μ_n (WU *et al.* 2001); for $r = 0$, $\hat{\mu}_{1,0}$ and $\hat{\mu}_{n,0}$ will be biased. To implement the test statistic (Eq. 1), we estimate σ^2 by $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{u}_i\right)^2$ and let $r = c\sigma_n$, where c > 0 is the penalty term. Cut-off values of $\Lambda_{n,r}$ are given in Table 1 in Wu et al. (2001). For example, for c = 0.1 and n = 100, the cut-off values of $\Lambda_{n,r}$ are 7.7 and 11.58 at the 5% and 1% levels, respectively. Simulations and several applications showed that such a test has very good power (WU *et al.* 2001). The magnitude of $\Lambda_{n,r}$ indicates the significance of the presence of a trend. Intuitively, the larger the $\Lambda_{n,r}$, the stronger the monotonic trend.

In practice, to test a decreasing trend in a series of observations, we simply flip the sign of the data and test the monotonic increasing trend. Because the spatial patterns studied here are very noisy, we applied this regression at different ranges to ensure that the significance of a trend is stable.

Super-sequences across codon positions

To examine the global spatial codon usage bias in a group of genes, we generated super-sequences for each codon position across all the genes in this group (Figure 1). We used the start

codon as the reference for super-sequences along the translational direction (forward) and the stop codon as the reference for super-sequences in the opposite direction (backward). We use forward and backward super-sequences to study the first half and the second half of genes separately. The modified effective number of codons ($\hat{N}_c'$) (NOVEMBRE 2002) for each super-sequence describes the overall codon usage bias at the codon position specified by the super-sequence. The statistic $\hat{N}_c'$ is a generalized form of the effective number of codons ($\hat{N}_c$) (WRIGHT 1990) and account for the background nucleotide composition. We use the super-sequence from the first and the second base to estimate the background nucleotide composition during the $\hat{N}_c'$ calculation. By studying the first half and the second half in opposite directions, this method is more sensitive to variations at codon positions than sliding-window methods. For comparison, we calculated both the original effective number of codons, $\hat{N}_c$, and the modified version, $\hat{N}_c'$.

When using this super-sequence method to analyze the intragenic spatial pattern in an entire genome, extreme length variation between genes can complicate the spatial pattern. We therefore group genes in a genome into equal intervals by gene length. Genes in each group have comparable lengths. The average length of each group divided by two is the length of the forward and backward super-sequences. The minimum number of genes in each interval (group) is 200. This ensures that each super-sequence by position has at least 200 codons. Simulation shows that the effective number of codons is more accurate as a proxy for codon usage bias when gene length is longer than 200 codons (WRIGHT 1990).

This super-sequence approach is also adopted to study the intragenic GC-content variation and the ratio of A/T-ending versus G/C ending major codons for Isoleucine, Leucine,

and Valanine in yeast genes. The A/T-ending major codons are: ATT, GTT, and GCT. The G/C-ending major codons are: ATC, GTC, and GCC (AKASHI 2003).

Implementation

Parsing of ORFs and sequence manipulations were coded in PERL. The modified effective number of codons ($\hat{N}_c'$) were calculated using the program ENCprime (NOVEMBRE 2002). The effective number of codons $\hat{N}_c$ were calculated using both ENCprime and the program CHIPS from the EMBOSS package (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/) (WRIGHT 1990; RICE *et al.* 2000). The R statistical package (http://www.r-project.org/) was used for statistical analysis and plots (RIPLEY 2001).

Key PERL and R scripts are available at the Li lab website (http://pondside.uchicago.edu/~lilab/research.html). All programs are only tested in a Red Hat Linux platform. The most recent versions of programs related to this work are also available upon request from H. Qin (hong_qin@urmc.rochester.edu).

RESULTS

Intragenic spatial codon usage bias in prokaryotic genomes

We chose to study three bacterial genomes, *Escherichia coli*, *Bacillus subtilis*, *Thermotoga maritima,* and one archaeal genome, *Sulfolobus tokodaii*. In each genome, we grouped genes into equal groups (bins) by gene length. This grouping is necessary because large gene length variation can obscure the global pattern generated by our super-sequence method. We chose $\hat{N}_c'$ as the measure of codon usage bias because it is insensitive to gene length

(WRIGHT 1990; COMERON and AGUADE 1998) and because it takes account of background nucleotide composition variation (NOVEMBRE 2002). $\hat{N}_c'$ is a generalized form of the effective number of codons, $\hat{N}_c$, which is a measure of the departure from equal usage of synonymous codons (WRIGHT 1990). A large $\hat{N}_c'$ or $\hat{N}_c$ value indicates weak bias, whereas a small value indicates strong bias. The maximal value of $\hat{N}_c'$ or $\hat{N}_c$ is 61, indicating equal usage of codons, i.e., no codon bias. For a gene with all the 20 kinds of amino acids and with the strongest codon usage bias, its $\hat{N}_c'$ or $\hat{N}_c$ value is 20.

We then generated super-sequences for each length group, both along and opposite to the translational direction, using the start and stop codons as references, respectively (start and stop codons are excluded in plots and isotonic regression). At each codon position, the first and second bases are chosen to generate a super-sequence to estimate the background nucleotide composition in the calculation of $\hat{N}_c'$. The $\hat{N}_c'$ of each super-sequence is plotted against its codon position (Figure 2). Because a small $\hat{N}_c'$ value indicates strong codon usage bias, we inverted the $\hat{N}_c'$ axis in the plot. Forward and backward plots are put side-by-side to represent the full-length spatial codon usage bias. For comparison, we always calculated both $\hat{N}_c'$ and $\hat{N}_c$ to check the consistency of spatial patterns.

The intragenic spatial codon usage bias in *E. coli* (Figure 2A and B) shows that codon usage bias gradually increases, plateaus, and then deceases sharply starting at ~50 codons from the stop codon. The pattern is consistent with previous reports in *E. coli* (BULMER 1988; EYRE-WALKER and BULMER 1993; EYRE-WALKER 1996a). The increasing trend of codon usage bias from ~25th codon up to the middle section is also visible in previous reports, although the

authors mainly drew attention to the initial regions (BULMER 1988; EYRE-WALKER and BULMER 1993).

To detect weak evolutionary forces in the $\hat{N}_c'$ versus codon position plot, which has large fluctuations, we applied an isotonic regression analysis to the first halves beyond the first 25 codons and the second halves before the last 25 codons. The adopted isotonic regression method uses a penalty factor to generate an unbiased test statistic, Λ (see Methods). The critical value of Λ is used to distinguish between a null hypothesis H_0: $\mu_1 = \mu_2 \ldots = \mu_n$ (constant) and the alternative hypothesis H_A: $\mu_1 \leq \mu_2 \ldots \leq \mu_n$ for an increasing trend.

We calculated Λ for the first half and the second half of the coding sequences, separately (Table 1). The critical value of the test statistic Λ at the 1% significance level is in the range of (12, 13) for sample sizes ranging from 100 to 1000 codons and penalty factor c=0.1. The gradual increasing trend from the 25th to the 200th codon is indeed significant at the 1% level for all five length groups of genes in *E. coli*. The gradual increasing and then plateau trend of codon usage bias can be intuitively understood from the plot of the regression mean (expected $\hat{N}_c'$) versus codon position (an example in Figure 3A and B). The plateau of this increasing trend is confirmed by the insignificant Λ values in the second halves and can be seen from the regression plot (Figure 3A and B).

Although the increasing trend of codon bias is consistent with purifying selection against nonsense errors, it may also be accounted for by changing selection pressure for mRNA structure after the initiation stage (EYRE-WALKER and BULMER 1993). To further explore the possible mechanism behind the gradually increasing codon usage bias from about the 25th to the 200th codon position, we chose to study the intragenic spatial patterns in *Bacillus subtilis* (Figure 2C and D), *Thermotoga maritima* (Figure 2E and F), and *Sulfolobus tokodaii* (Figure 2G and H).

In *B. subtilis*, codon usage bias decreases near the start, but gradually increases along the translational direction, and then plateaus toward the end. In *T. maritima*, codon usage bias gradually increases up to the 150th codon position, plateaus, and then drops near the 3' end. The codon usage bias in *S. tokodaii* follows a pattern similar to that in *T. maritima*. Codon usage bias pattern is also variable near the initiation regions in both *T. maritima* and *S. tokodaii*. Thus, in the four prokaryotic genomes examined, the spatial pattern is variable near the initiation region, but consistently increases gradually from about the 25th to the 200th codon. Therefore, different evolutionary forces must predominate in the initial region and in the central segments of genes, respectively. We argue that the gradual increasing codon usage bias is due to stronger purifying selection to eliminate errors with higher costs in late elongation steps (BULMER 1988; KURLAND 1992; EYRE-WALKER and BULMER 1993).

Both $\hat{N}_c'$ and $\hat{N}_c$ give similar spatial patterns for the four prokaryotic genomes examined. The absolute values of $\hat{N}_c'$ can differ by as large as 10 from the values of $\hat{N}_c$ in *S. tokodaii*, yet their spatial patterns are largely unchanged, as supported by results from isotonic regressions (Table 1). The differences between $\hat{N}_c'$ and $\hat{N}_c$ values does affect the significance evaluation, because large differences are usually more easily detected than small ones for a given sample size. Because $\hat{N}_c'$ takes nucleotide composition into account, we use Λ from $\hat{N}_c'$ values to evaluate the significance of a bias trend.

Noticeably, in both *E. coli* and *B. subtilis*, the gradual increasing trends are more conspicuous in longer genes than in shorter genes.

Intragenic spatial codon usage bias in *S. cerevisiae* genome

The spatial codon usage bias pattern in *S. cerevisiae* resembles that of prokaryotic genomes, but the gradual increasing trend is discernable even in the second half of yeast genes (Figure 4A and B). In Lc≈300, 400, and 515 length intervals (Lc is the number of codons in a gene), the $\hat{N}_c'$ plot drops near the start codon, increases along the translational direction from about the 25th to the 200th codons, plateaus for about 100 codons, and then gradually increases again up to the -25th codons, and finally drops near the end (Figure 3C and D, Figure 4A and B). The increasing trend in the second half is consistent in genes shorter than 600 codons and is significant by isotonic regression (Table 1 and Figure 3C and D). The increasing trend from the +25th to the +400th is also significant in the Lc≈1175 interval, but the second half shows an opposite trend as compared to shorter length intervals.

The $\hat{N}_c$ analysis in yeast shows a similar increasing trend of bias along translational direction from the +25th to the +400th codon (plots are not shown, but isotonic regression results are given in Table 1). The $\hat{N}_c'$ plots are usually noisier than the $\hat{N}_c$ plots, maybe because more parameters have to be estimated in the $\hat{N}_c'$ plots than in the $\hat{N}_c$ plots.

We then studied the effect of gene expression on the spatial pattern (Figure 4C and D, Table 2). Holstege et al. (1998) estimated the number of mRNA molecules of each yeast gene in a single wild-type haploid cell using high-density oligonucleotide arrays. The estimation is conducted for cells grown to the mid-log phase in rich media and has been shown to be informative in codon usage analysis (AKASHI 2003). We picked genes from the top 20%, the middle 20%, and the bottom 20% by the transcript abundance and compared their spatial bias patterns. There is clearly a shift of the spatial patterns between the top and middle 20% expression levels. The spatial patterns are mostly flat in the bottom 20% expression level, indicating that intragenic spatial codon usage bias is dependent upon expression level. The

increasing trends in short genes are comparable in top and middle expression levels, indicating that changes of spatial patterns can be de-coupled from changes of expression levels.

Because non-selective forces are often correlated with changes of GC content (MARAIS 2003), we investigated the intragenic spatial pattern of GC content using the first and second base of each codon position. Although $\hat{N}_c'$ has taken the background nucleotide composition into account, investigation of non-selective forces can help us to better understand the observed incremental pattern of codon bias. The intragenic spatial pattern of GC content variation is also dependent upon expression levels (Figure 5A and B). The spatial pattern appear to be U-shape in lowly expressed genes, but the U-shape is less conspicuous in highly expressed genes, which is supported by isotonic regression (Table 3). The overall GC content is higher in highly expressed genes than in lowly expressed ones. Noticeably, the U-shaped spatial pattern of GC content variation is more or less symmetric, whereas the incremental shape of the codon bias is asymmetric.

To further test the role of GC-content variation, we calculated the ratio of A/T-ending major codon number versus G/C-ending major codon number for Isoleucine, Leucine, and Alanine. In yeast, these three codon families contain two major codons, one with A/T-ending and another with G/C-ending (AKASHI 2003). We calculated the intragenic spatial distribution of this ratio in genes grouped by expression levels. The spatial pattern of this ratio appears to be flat in all groups of genes (data not shown). When applying isotonic regression at different ranges, no stable significant regression score was found.

Intragenic spatial codon usage bias in *D. melanogaster* genome

D. melanogaster is of particular interest to us because its codon usage bias was suggested to be nonunimformly influenced by the Hill-Robertson effect (COMERON and KREITMAN 2002). Indeed, *D. melanogaster* reveals a distinctive M-shaped global spatial codon usage bias (Figure 6A and B), with less bias in the middle and both ends. The low codon usage bias in the middle is statistically significant, confirmed by the decreasing trend in the first halves and the increasing trend in the second halves of genes shown by isotonic regression in the central segments (Table 1). The plots of longer genes are often below the plots of shorter gene, especially in the middle section, indicating that overall codon usage bias of longer genes tends to be weaker than that of shorter genes. The $\hat{N}_c$ analysis in *D. melanogaster* gives the same M-shaped global spatial pattern as does the $\hat{N}_c'$ analysis (plots are not shown, but isotonic regression results are given in Table 1).

Next, we investigated the effect of expression level on the M-shaped spatial pattern. Montalta-He et al. (2002) provided four independent measurements of gene expression level in wild-type *D. melanogaster* using high-density oligonucleotide arrays. We used the averaged intensity value of these four independent hybridizations as our estimate of expression level for each gene. We obtained expression measurements from Montalta-He et al.'s experiments for over 7300 single-isoform ORFs (Release 3.1). We then picked genes from the top 20% and bottom 20% of the distribution at expression levels and compared their spatial codon usage bias (Figure 6C and D). The plots are noisier than the whole genome plots due to the smaller data sets here. Nevertheless, expression level only shifts the spatial bias pattern without dramatically altering the M-shape.

We then revealed the intragenic GC content variation using the first and second base of each codon position (Figure 7A and B). The overall pattern of intragenic GC content variation is

also M-shape, however, the sharp variations are located mostly within the first and last 50 bp. In contrast, the pattern of codon bias peaks around +100th bp and -100th bp (Figure 6). In addition, variation of nucleotide composition at the first and second bases is corrected in the calculation of $\hat{N}_c'$.

The presence of introns is expected to mitigate the effect of interference (COMERON and KREITMAN 2002), so we compared the spatial patterns $\hat{N}_c'$ in intronless and two-exoned genes. In long genes (average length greater than 500 ~ 600 bp), the M-shaped spatial pattern of intronless genes appear to be more conspicuous than that of two-exoned ones, in the sense that the latter is more flat (Figure 8A and B). In the middle section (from +200th to -200th), $\hat{N}_c'$ is significantly higher in intronless genes than in two-exoned genes (t-test, p-value = 0.01), indicating the significantly lower codon bias in the middle section of intronless genes. For comparison, we calculated the intragenic pattern of GC content variation in the two groups of genes (Figure 8C and D). In the middle section, GC variation between the two gene groups is not significantly different by t-test.

DISCUSSION

Variation of codon usage bias near the start and stop codons

In most genomes, we observed sharp variations near the start and stop codons. In addition to the weakened elongation-related selection (EYRE-WALKER and BULMER 1993), several other possibilities may account for the variation of codon bias near the start and stop codons. First, the ends of ORFs are more likely to be mis-annotated, especially in the yeast and fruit fly genomes. We hope to have mitigated this possibility in yeast to some extent by excluding ORFs with

uncertain annotations. A recent comparative analysis of several yeast genomes proposed to change the annotations of ~500 yeast ORFS (KELLIS *et al.* 2003). For simplicity, we excluded these questioned ORFs from the downloaded *S. cerevisiae* genome. The starting ATG codons in the prokaryotic genomes can be identified with high confidence by the presence of an upstream Shine-Dalgarno sequence (SAITO and TOMITA 1999; SAKAI *et al.* 2001). Therefore, it is unlikely that the codon usage bias variation near the initiation ATG in the prokaryotic genomes is due to mis-annotations. There is evidence that the occurrence of ATG in the 5' UTR is highly suppressed in the eukaryotic genomes (SAITO and TOMITA 1999), which indicates that the initiation codon ATG is less like to be mis-annotated than the stop codon.

Second, the ends of ORFs are evolutionarily more variable than the middle section of the coding sequence. The possibility is supported by the comparison of several yeast genomes (KELLIS *et al.* 2003). The *S. cerevisiae* genome contains about 250 ORFs that show variable stop codon positions in other *sensu stricto* yeast genomes, suggesting that coding regions near the stop codon are indeed evolutionarily variable (KELLIS *et al.* 2003).

Third, the base composition near both ends is different from that in the middle section of the reading frame. The modified effective number of codons, $\hat{N}_c'$, takes the background nucleotide composition into account. In practice, it is hard to find a "perfect" background sequence. We estimate the background nucleotide composition by using the first and second bases in a given codon position to generate super-sequences. This may not be a good choice, because nonsynonymous sites show a weaker correlation with GC content than synonymous sites in prokaryotic genomes (LOBRY 1997; SINGER and HICKEY 2000). In *E. coli*, base composition is variable in the first 15-20 codons, but this variation disappears after the first 20 codons (EYRE-WALKER and BULMER 1993). However, the consistent trend of both $\hat{N}_c'$ and $\hat{N}_c$ after the initial

region in four prokaryotic genomes begs a common explanation other than base composition variation. Further analysis will be needed to better understand the nature of the codon usage bias variation in regions near the start and stop codons.

The incremental pattern of intragenic spatial codon usage bias in yeast and prokaryotic genomes

Both non-selective forces and selective forces have been argued to shape the codon usage bias in yeast (for reviews see AKASHI 2001 and MARAIS 2003). Here, we show that codon usage bias and GC content variation follow characteristically different spatial patterns within yeast genes. The incremental spatial pattern of codon usage bias is asymmetric and is stronger in highly expressed genes (Figure 4C and D). The spatial pattern of GC content variation is more or less a symmetric U-shape and is stronger in lowly expressed genes (Figure 5A and B).

Various arguments can explain this incremental pattern of codon usage bias. One argument is the 5'-3' polarity in gene conversion (SCHULTES and SZOSTAK 1990; NICOLAS and PETES 1994; NICOLAS 1998). This polarity in conjunction with biased repair toward GC (MARAIS 2003) would then explain the incremental spatial pattern of codon bias in yeast. The intragenic spatial pattern of GC content variation shows a moderate feature of polarity, but its overall pattern is largely symmetric. This pattern indeed suggests the role of recombination, but its characteristics are in discordance with the incremental nature of the spatial codon pattern in yeast. Using the ratio of A/T-ending versus G/C-ending major codon numbers of Isoleucine, Leucine, and Alanine, we did not detect significant influence of GC content variance on intragenic codon bias, although this negative result may be due to the weak power of this method. Furthermore, because GC content variations are also influenced by gene expression

levels in yeast, association with GC content variation is not a good indicator for either selective or non-selective forces in this organism. It is also not obvious why biased gene conversion ought to be correlated with gene expression levels. Finally, nucleotide composition variation is corrected in the calculation of pattern of codon usage bias. Hence, the strength of codon bias variation is much stronger than the expected variation exclusively contributed by nucleotide composition variation.

Selection against nonsense errors is another argument to explain this kind of position dependent effect, because costs for errors are higher in late elongation steps (KURLAND 1992; AKASHI 2001). This straightforward argument is favored by us on ground of parsimony. Both selection against mis-incorporation and selection for translation efficiency are expected to be position independent (or gene length-independent), and have to be ruled out.

In *E. coli* and *B. subtilis*, the incremental pattern is stronger in longer genes than in shorter genes. It is known that codon usage bias is positively correlated with gene length in *E. coli* (EYRE-WALKER 1996b; MORIYAMA and POWELL 1998). Stronger spatial bias in longer genes would then explain the observed positive correlation between codon usage bias and gene length (EYRE-WALKER 1996b; MORIYAMA and POWELL 1998). In other words, the biased spatial pattern reflects the overall biased usage of synonymous codons in these genomes. Interestingly, the incremental pattern in *E. coli* was attributed to non-selective factors (EYRE-WALKER and BULMER 1993; EYRE-WALKER 1996a), but positive association between gene length and codon bias was proposed to be due to selection against nonsense errors (EYRE-WALKER 1996a). We argue that purifying selection against nonsense errors is a consistent explanation for both the incremental spatial pattern and the association between codon usage bias and gene length.

Although purifying selection against nonsense errors is our best argument so far, we do not exclude the possibilities of other factors, such as constraints at the protein conformation level.

The M-shaped intragenic spatial codon usage bias in *D. melanogaster* genome

The M-shaped intragenic pattern of codon usage bias has taken into account of the M-shaped GC content variation. The spatial pattern of codon bias is influenced by expression levels and peaks around +100th codon position in the 5' region and broadly around -100th in the 3' region (Figure 6C and D), whereas the spatial pattern of GC content variation appears to be insensitive to changes of expression levels and peaks +50th and -50th codon positions (Figure 7A and B). The presence of introns significantly influences the pattern of codon usage bias but not the pattern of GC-content variation.

In *D. melanogaster*, GC-biased mutation and/or GC-biased gene-conversion has been postulated to contribute to codon usage bias differences between genes, i.e., the intergenic pattern (MARAIS et al. 2001; MARAIS 2003; MARAIS et al. 2003; but see also HEY and KLIMAN 2002; BETANCOURT and PRESGRAVES 2002). Based on this argument, the M-shaped pattern of GC content variation would suggest intragenic variation of recombination rate. Hence, the spatial pattern of codon usage bias reflects the pattern of recombination rate variation. However, this argument would ignore the correction of GC background in the calculation of codon usage bias. In addition, the spatial pattern of GC-content variation is characteristically distinct from the spatial pattern of codon usage bias.

Under the Hill-Robertson interference model, Comeron and Kreitman proposed that more sites are linked in the middle section of the gene than other regions of the gene, thereby resulting

in stronger interference in the middle section (COMERON and KREITMAN 2002). This argument is more straightforward than the GC-content version and is further supported by the comparison between intronless and two-exoned genes. It should be noted that the Hill-Robertson effect alone should give a U-shaped pattern. The M-shape hence indicates the presence of other factors. It is interesting that the M-shaped spatial bias peaks around +100th (forward) and around -100th codon positions (backward). This may be informative about the interaction of the weak selection for codon usage bias, mutation rate, and recombination rate.

Besides the assumed weak selection on synonymous codon usage, there is the alternative argument of strong selection. As Fay and Wu showed that strong selection and hitch-hiking can occur in a ~350 bp region, short enough to influence intragenic distribution of codon usage bias (FAY and WU 2000). The key difference between the strong and weak selection models is the number of sites involved in the formation of the M-shape. Strong selection models typically imply selection at a single site, whereas weak selection and interference models require multiple sites. We think that the gradual decreasing trend toward the middle section is more consistent with weak selection at multiple sites.

As with yeast, it is very difficult to distinguish selective versus non-selective forces in fruit fly genes (DURET 2002). Although we think that the Hill-Robertson interference is the most parsimonious argument, the question is still open and more studies are needed. A genome-scale comparison of gene structures and intragenic spatial patterns of codon usage bias between *D. melanogaster* and *D. pseudoobscura* may be informative for this subject.

Why lack of interference in yeast and prokaryotic genes?

It is perplexing that there is no detectable signature of the Hill-Robertson effect in yeast and its spatial codon usage bias resembles that of prokaryotes. Although recombination during sexual reproduction occurs at a much higher rate in the yeast genome than in the fruit fly genome in laboratory (CHERRY *et al.* 1997; GERTON *et al.* 2000), sexual reproduction of yeast may rarely occur in nature (JENSEN *et al.* 2001; WINZELER *et al.* 2003). Simulation showed that the intragenic spatial pattern of codon usage bias is insignificant when recombination is low (COMERON and KREITMAN 2002; COMERON unpublished data). At the extreme case of no recombination, the Hill-Robertson effect will be the same at every position of the sequence and will thus produce no signature on the spatial pattern of codon usage bias. Hence, a very low effective recombination rate may account for the absence of a signature of the Hill-Robertson effect on the spatial pattern of codon usage bias in yeast and prokaryotes.

On the other hand, the effective recombination rate of yeast may be indeed much higher in yeast than in fruit fly, which can also explain the discrepancy between these two species. The uncertainty on this issue highlights the importance to study the polymorphic variations in natural isolates of yeast.

The effect of expression levels on intragenic spatial codon usage bias

The positive correlation between codon usage bias and expression levels are often attributed to selection for translational efficiency (GOUY and GAUTIER 1982; SHARP and LI 1986; DURET and MOUCHIROUD 1999; AKASHI 2001; CARLINI and STEPHAN 2003). Selection against nonsense errors is largely suggested by the positive correlation between codon usage bias and gene length in some genomes (MORIYAMA and POWELL 1998; COGHLAN and WOLFE 2000). Selection against mis-sense errors is based on the positive correlation between codon usage bias

and conservation of amino acids (AKASHI 1994). The selection effect of translational efficiency and against mis-sense errors on the spatial pattern of codon bias is expected to be uniform, whereas the selection effect against nonsense errors is expected to be incremental along the translational direction, i.e., asymmetric. Hence, studying the intragenic spatial pattern of codon bias can dissect the roles of these evolutionary forces. If an asymmetric spatial pattern is observed, it is unlikely to be the result of a process with a uniform spatial effect.

By choosing appropriate expression levels, we show that expression levels significantly affect the overall codon usage bias in both yeast and fruit fly genes, but to a less extent on the intragenic spatial pattern of codon usage bias. Therefore, we argue that selection for translational efficiency significantly affects the overall codon usage bias, i.e., the average, whereas purifying selection against nonsense errors and the Hill-Robertson effect greatly influences the shape of the spatial codon usage bias, i.e., the spatial distribution.

Final remarks

Compared with some previous methods in analyzing patterns of codon usage bias (BULMER 1988; HARTL *et al.* 1994; EYRE-WALKER 1996b; COMERON *et al.* 1999; COMERON and KREITMAN 2002; KLIMAN *et al.* 2003), our method is more sensitive for addressing variations by positions and identifying the trend of changes in spatial patterns. This enables us to distinguish the subtle signatures of various evolutionary forces. The gradually increasing spatial bias pattern in *S. cerevisiae* and prokaryotic genomes suggests purifying selection against nonsense errors. The M-shaped global spatial codon usage bias confirms the low codon usage bias in the middle of genes in *D. melanogaster*, which may be best explained by the interference of weak selections (COMERON and KREITMAN 2002). Selection for translational efficiency mainly affects the overall

codon usage bias, whereas selection against nonsense errors and the Hill-Robertson interference leave discernable signatures in the intragenic spatial pattern. Therefore, our analysis suggests two types of selective process on codon bias: One is the optimization of the biochemical process of translation, and the other is the subtle process due to evolutionary mechanisms and is best understood from theory of population genetics. The first is common to most species. The second, acting upon the first type of selection, is expected to differ among species because it is sensitive to the effective population size and the evolutionary history of each species. Hence, further analysis of global spatial codon usage bias in other genomes may help us better understand the evolution of codon usage in those species.

ACKNOWLEDGMENTS

We thank Hiroshi Akashi, an anonymous reviewer, and Jianming Zhang for valuable suggestions. We are grateful to Manyuan Long, Robert Friedman, Todd Oakley, Peter Bauman, Zheng-yuan Zhu, and many others for helpful discussions. We also thank Michael Stein, Mei Wang, Stacy Steinberg, Jing Yang, and George Kordzakhia in the Consulting Program at the Department of Statistics, University of Chicago for their helpful suggestions. This study was supported by grant GM30998 from NIH to WH Li.

LITERATURE CITED

AKASHI, H., 1994 Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics **136:** 927-935.

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067-1076.

AKASHI, H., 1997 Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene **205:** 269-278.

AKASHI, H., 2001 Gene expression and molecular evolution. Curr Opin Genet Dev **11:** 660-666.

AKASHI, H., 2003 Translational selection and yeast proteome evolution. Genetics **164:** 1291-1303.

AKASHI, H., R. M. KLIMAN and A. EYRE-WALKER, 1998 Mutation pressure, natural selection, and the evolution of base composition in Drosophila. Genetica **102-103:** 49-60.

AKASHI, H., and S. W. SCHAEFFER, 1997 Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics **146:** 295-307.

ANTEZANA, M. A., and M. KREITMAN, 1999 The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J Mol Evol **49:** 36-43.

BEGUN, D. J., 2001 The frequency distribution of nucleotide variation in Drosophila simulans. Mol Biol Evol **18:** 1343-1352.

BERG, O. G., 1996 Selection intensity for codon bias and the effective population size of Escherichia coli. Genetics **142:** 1379-1382.

BETANCOURT, A. J., and D. C. PRESGRAVES, 2002 Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci U S A **99:** 13616-13620.

BIRDSELL, J. A., 2002 Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol **19:** 1181-1197.

BULMER, M., 1987 Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728-730.

BULMER, M., 1988 Codon usage and intragenic position. J Theor Biol **133:** 67-71.

BULMER, M., 1990 The effect of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res **18:** 2869-2873.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897-907.

BURNS, D. M., and I. R. BEACHAM, 1985 Rare codons in E. coli and S. typhimurium signal sequences. FEBS Lett **189:** 318-324.

CARLINI, D. B., Y. CHEN and W. STEPHAN, 2001 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 **159:** 623-633.

CARLINI, D. B., and W. STEPHAN, 2003 In Vivo Introduction of Unpreferred Synonymous Codons Into the Drosophila Adh Gene Results in Reduced Levels of ADH Protein. Genetics **163:** 239-243.

CHEN, G. F., and M. INOUYE, 1990 Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. Nucleic Acids Res **18:** 1465-1473.

CHEN, G. T., and M. INOUYE, 1994 Role of the AGA/AGG codons, the rarest codons in global gene expression in Escherichia coli. Genes Dev **8:** 2641-2652.

CHEN, Y., D. B. CARLINI, J. F. BAINES, J. PARSCH, J. M. BRAVERMAN *et al.*, 1999 RNA secondary structure and compensatory evolution. Genes Genet Syst **74:** 271-286.

CHERRY, J. M., C. BALL, S. WENG, G. JUVIK, R. SCHMIDT *et al.*, 1997 Genetic and physical maps of Saccharomyces cerevisiae. Nature **387:** 67-73.

COGHLAN, A., and K. H. WOLFE, 2000 Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast **16:** 1131-1145.

COMERON, J. M., unpublished data.

COMERON, J. M., and M. AGUADE, 1998 An evaluation of measures of synonymous codon usage bias. J Mol Evol **47:** 268-274.

COMERON, J. M., and M. KREITMAN, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 389-410.

COMERON, J. M., M. KREITMAN and M. AGUADE, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239-249.

DURET, L., 2002 Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev **12:** 640-649.
DURET, L., and D. MOUCHIROUD, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A **96:** 4482-4487.
EYRE-WALKER, A., 1996a The close proximity of Escherichia coli genes: consequences for stop codon and synonymous codon use. J Mol Evol **42:** 73-78.
EYRE-WALKER, A., 1996b Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Mol Biol Evol **13:** 864-872.
EYRE-WALKER, A., and M. BULMER, 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res **21:** 4599-4603.
EYRE-WALKER, A., and M. BULMER, 1995 Synonymous substitution rates in enterobacteria. Genetics **140:** 1407-1412.
FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.
FITCH, W. M., 1980 Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNA's. J Mol Evol **16:** 153-209.
GALTIER, N., 2003 Gene conversion drives GC content evolution in mammalian histones. Trends Genet **19:** 65-68.
GALTIER, N., G. PIGANEAU, D. MOUCHIROUD and L. DURET, 2001 GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159:** 907-911.
GERTON, J. L., J. DERISI, R. SHROFF, M. LICHTEN, P. O. BROWN *et al.*, 2000 global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A **97:** 11383-11390.
GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res **10:** 7055-7074.
HARTL, D. L., E. N. MORIYAMA and S. A. SAWYER, 1994 Selection intensity for codon bias. Genetics **138:** 227-234.
HEY, J., and R. M. KLIMAN, 2002 Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics **160:** 595-608.
HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet Res **8:** 269-294.
HOLSTEGE, F. C., E. G. JENNINGS, J. J. WYRICK, T. I. LEE, C. J. HENGARTNER *et al.*, 1998 Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95:** 717-728.
IKEMURA, T., 1981 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 **151:** 389-409.
IKEMURA, T., 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol **158:** 573-597.
IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol **2:** 13-34.
JENSEN, M. A., H. L. TRUE, Y. O. CHERNOFF and S. LINDQUIST, 2001 Molecular population genetics and evolution of a prion-like protein in Saccharomyces cerevisiae. Genetics **159:** 527-535.
KELLIS, M., N. PATTERSON, M. ENDRIZZI, B. BIRREN and E. S. LANDER, 2003 Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:** 241-254.
KLIMAN, R. M., 1999 Recent selection on synonymous codon usage in Drosophila. J Mol Evol **49:** 343-351.
KLIMAN, R. M., and A. EYRE-WALKER, 1998 Patterns of base composition within the genes of Drosophila melanogaster. J Mol Evol **46:** 534-541.
KLIMAN, R. M., and J. HEY, 1993 Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol Biol Evol **10:** 1239-1258.
KLIMAN, R. M., and J. HEY, 1994 The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics **137:** 1049-1056.
KLIMAN, R. M., and J. HEY, 2003 Hill-Robertson interference in Drosophila melanogaster: reply to Marais, Mouchiroud and Duret. Genet Res **81:** 89-90.
KLIMAN, R. M., N. IRVING and M. SANTIAGO, 2003 Selection conflicts, gene expression, and codon usage trends in yeast. J Mol Evol **57:** 98-109.
KURLAND, C. G., 1992 Translational accuracy and the fitness of bacteria. Annu Rev Genet **26:** 29-50.
LI, W. H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol **24:** 337-345.

LOBRY, J. R., 1997 Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene **205:** 309-316.
MARAIS, G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends Genet **19:** 330-338.
MARAIS, G., and B. CHARLESWORTH, 2003 Genome evolution: recombination speeds up adaptive evolution. Curr Biol **13:** R68-70.
MARAIS, G., D. MOUCHIROUD and L. DURET, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A **98:** 5688-5692.
MARAIS, G., D. MOUCHIROUD and L. DURET, 2003 Neutral effect of recombination on base composition in Drosophila. Genet Res **81:** 79-87.
MARAIS, G., and G. PIGANEAU, 2002 Hill-Robertson interference is a minor determinant of variations in codon bias across Drosophila melanogaster and Caenorhabditis elegans genomes. Mol Biol Evol **19:** 1399-1406.
MASIDE, X., A. W. LEE and B. CHARLESWORTH, 2004 Selection on codon usage in Drosophila americana. Curr Biol **14:** 150-154.
MCVEAN, G. A., and B. CHARLESWORTH, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929-944.
MCVEAN, G. A., and G. D. HURST, 2000 Evolutionary lability of context-dependent codon bias in bacteria. J Mol Evol **50:** 264-275.
MODIANO, G., G. BATTISTUZZI and A. G. MOTULSKY, 1981 Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? Proc Natl Acad Sci U S A **78:** 1110-1114.
MONTALTA-HE, H., R. LEEMANS, T. LOOP, M. STRAHM, U. CERTA *et al.*, 2002 Evolutionary conservation of otd/Otx2 transcription factor action: a genome-wide microarray analysis in Drosophila. Genome Biol **3:** RESEARCH0015.
MORIYAMA, E. N., and J. R. POWELL, 1997 Codon usage bias and tRNA abundance in Drosophila. J Mol Evol **45:** 514-523.
MORIYAMA, E. N., and J. R. POWELL, 1998 Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res **26:** 3188-3193.
NICOLAS, A., 1998 Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. Proc Natl Acad Sci U S A **95:** 87-89.
NICOLAS, A., and T. D. PETES, 1994 Polarity of meiotic gene conversion in fungi: contrasting views. Experientia **50:** 242-252.
NIIMURA, Y., M. TERABE, T. GOJOBORI and K. MIURA, 2003 Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. Nucleic Acids Res **31:** 5195-5201.
NOVEMBRE, J. A., 2002 Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol **19:** 1390-1394.
OHNO, H., H. SAKAI, T. WASHIO and M. TOMITA, 2001 Preferential usage of some minor codons in bacteria. Gene **276:** 107-115.
PLOTKIN, J. B., and J. DUSHOFF, 2003 Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc Natl Acad Sci U S A **100:** 7152-7157.
RAND, D. M., and L. M. KANN, 1998 Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. Genetica **102-103:** 393-407.
RICE, P., I. LONGDEN and A. BLEASBY, 2000 EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16:** 276-277.
RIPLEY, B., 2001 The R project in statistical computing. MSOR Coneections **1:** 23-25.
ROBERTSON, T., F. T. WRIGHT and R. L. DYKSTRA, 1988 *Order Restricted Statistical Inference*. Wiley & sons., New York.
SAITO, R., and M. TOMITA, 1999 On negative selection against ATG triplets near start codons in eukaryotic and prokaryotic genomes. J Mol Evol **48:** 213-217.
SAKAI, H., C. IMAMURA, Y. OSADA, R. SAITO, T. WASHIO *et al.*, 2001 Correlation between Shine--Dalgarno sequence conservation and codon usage of bacterial genes. J Mol Evol **52:** 164-170.
SCHULTES, N. P., and J. W. SZOSTAK, 1990 Decreasing gradients of gene conversion on both sides of the initiation site for meiotic recombination at the ARG4 locus in yeast. Genetics **126:** 813-822.
SHARP, P. M., and W. H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol **24:** 28-38.
SHARP, P. M., M. STENICO, J. F. PEDEN and A. T. LLOYD, 1993 Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans **21:** 835-841.

SHIELDS, D. C., and P. M. SHARP, 1987 Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. Nucleic Acids Res **15:** 8023-8040.

SINGER, G. A., and D. A. HICKEY, 2000 Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol **17:** 1581-1588.

SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry*. W.H. Freeman and Company, New York.

STENSTROM, C. M., H. JIN, L. L. MAJOR, W. P. TATE and L. A. ISAKSSON, 2001 Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in Escherichia coli. Gene **263:** 273-284.

WINZELER, E. A., C. I. CASTILLO-DAVIS, G. OSHIRO, D. LIANG, D. R. RICHARDS *et al.*, 2003 Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. Genetics **163:** 79-89.

WRIGHT, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23-29.

WU, W. B., M. WOODROOFE and G. MENTZ, 2001 Isotonic regression: Another look at the changepoint problem. Biometrika **88:** 793-804.

Table 1. Monotonic trend of intragenic spatial codon usage bias along the translational direction

Species	Lc	First half			Second half		
		λ of $\hat{N}_c$	λ of $\hat{N}_c$'	Codon bias trend along translational direction	λ of $\hat{N}_c$	Λ of $\hat{N}_c$'	Codon bias trend along translational direction
E. coli	185	21.8***	19.8***	Increase	0.6	0.1	Plateau
[+25,-25]	255	22.6***	23.5***	Increase	1.5	1.3	Plateau
	325	58.3***	41.7***	Increase	14.1***	15.4***	Plateau, decrease
	420	76.1***	69.4***	Increase	7.8*	7.7*	Plateau, decrease
	675	243.8***	243.5***	Increase	34.1***	31.3***	Plateau, decrease
B. subtilis	175	1.7	8.9**	Increase	8.2**	3.8	Plateau
[+25,-25]	230	0.0	0.3	Plateau	0.1	0.0	Plateau
	285	9.5**	9.3**	Increase	4.9	5.3	Plateau
	345	0.1	2.3	Plateau	1.4	0.1	Plateau
	430	2.1	9.7**	Increase	0.2	1.2	Plateau
	710	36.9***	49.0***	Increase	4.8	8.2*	Decrease
T. maritima	200	18.4***	16.4***	Increase	8.9**	6.9	Plateau
[+25,-25]	285	3.9	8.8**	Increase	0.2	3.1	Plateau
	375	7.0	5.0	Plateau	11.1**	12.2***	Decrease
	610	30.7***	24.7***	Increase	1.1	0.4	Plateau
S. tokodaii	185	5.1	10.0**	Increase	6.5	9.0**	Increase, decrease
[+25,-25]	255	0.6	0.0	Plateau	0.0	1.1	Plateau
	350	0.1	7.2*	Increase	1.8	3.1	Plateau
	570	7.8*	8.2**	Increase	4.2	12.1**	Plateau, decrease
Yeast	205	2.4	3.4	Plateau	12.6***	21.8***	Increase
[+25,-25]	300	23.2***	22.1***	Increase	12.7***	25.3***	Increase
	400	32.5***	8.1*	Increase	4.9	22.3***	Increase
	515	46.0***	8.5**	Increase	2.7	12.2***	Increase
	680	114.3***	1.8	Plateau	0.0 (12.7***)	4.6(1.0)	Plateau
	1175	289.6***	20.3***	Increase	0.0(68.5***)	0.1(43.8***)	(Decrease)
Fruit fly	180	NA	NA	NA	NA	NA	NA
[+110,-110]	240	5.7	1.0	Plateau	2.2	1.1	Plateau
	305	5.8	8.4***	Decrease	4.4	8.1***	Increase
	380	15.4***	14.6***	Decrease	26.3***	27.0***	Increase
	460	22.9***	13.3***	Decrease	3.1	2.5	Plateau
	565	1.5	1.0	Plateau	4.9	0.5	Plateau
	745	87.9***	55.4***	Decrease	10.1***	2.3	Plateau
	1515	262.7***	135.6***	Decrease	126.8***	39.3***	Increase

Significance at 0.01 level (***), 0.05 level (**), and nearly 0.05 (*). λ : The test statistic of isotonic regression. In parenthesis are the λ values and spatial bias patterns for the opposite trends, because isotonic regression detects only a monotonic trend. Lc: Average gene length measured in the number of codons. Genes are grouped into equal intervals by their lengths. In

parentheses are the Λ values calculated for the first halves and second halves excluding only the start and stop codons. In brackets are the ranges in which isotonic regression is applied. The bias trend of the significant patterns are interpreted based on only the $\hat{N}_c'$ results, although $\hat{N}_c$ and $\hat{N}_c'$ often give similar interpretations. Additionally, when the increasing or decreasing pattern is significant in the first half, the non-significant pattern in the second half is interpreted as "plateau". When a regression plot shows that significant changes in second half are only near the 3' end, the trends are interpreted as "plateau, increasing" or "plateau, decreasing" accordingly. In Lc=185 interval for *S. tokodaii*, the trend in the second half increases and then decreases. Penalty factor c=0.1 for isotonic regression.

Table 2. Effect of gene expression level on the intragenic spatial pattern of codon bias in yeast

Expression level	Lc	First half			Second half		
		λ of $\hat{N}_c$	λ of $\hat{N}_c$'	Codon bias trend along translational direction	λ of $\hat{N}_c$	λ of $\hat{N}_c$'	Codon bias trend along translational direction
Top 20%	215	3.3	10.9**	Increase	15.4***	26.5***	Increase
	375	72.3***	35.9***	Increase	2.9	14.3***	Increase
	795	83.4***	32.5***	Increase	1.2	4.0	Plateau
Middle 20%	275	9.5	10.4**	Increase	9.9**	24.5***	Increase
	485	10.9**	4.0	Plateau	7.8	14.8***	Increase
	980	39.7***	1.4	Plateau	0.8	6.5	Plateau
Bottom 20%	270	2.7	1.4	Plateau	4.2	2.4	Plateau
	500	8.2	1.1	Plateau	3.2	6.1	Plateau
	930	74.1***	4.6	Plateau	0.4	0.6	Plateau

Significance at 0.01 level (***) and 0.05 level (**). λ : The test statistic of isotonic regression. Lc: Average gene length measured in the number of codons. In each expression level, genes are grouped into equal intervals by their lengths. The trend is analyzed after the first 25 codons and before the last 25 codons. The bias trend of the significant patterns is interpreted based on only the $\hat{N}_c$' results. Penalty factor c=0.1 for isotonic regression.

Table 3. Intragenic spatial pattern of GC content in *S. cerevisiae*.

Expression level	Lc	First half		Second half	
		λ of *GC%*	Trend along translational direction	λ of *GC%*	Trend along translational direction
Top 20%	215	12.1***	Decrease	4.3	Plateau
	375	6.7	Plateau	6.2	Plateau
	795	11.1**	Decrease	9.3**	Increase
Middle 20%	275	3.5	Plateau	3.9	Plateau
	485	12.3***	Decrease	7.9	Plateau
	980	83.1***	Decrease	27.1***	Increase
Bottom 20%	270	9.1**	Decrease	6.5	Plateau
	500	35.9***	Decrease	6.2	Plateau
	930	108.5***	Decrease	20.2***	Increase

Significance at 0.01 level (***) and 0.05 level (**). λ : The test statistic of isotonic regression. Lc: Average gene length measured in the number of codons. In each expression level, genes are grouped into equal intervals by their lengths. The trend is analyzed after the first 25 codons and before the last 25 codons. Penalty factor c=0.1 for isotonic regression.

FIGURE LEGEND

Figure 1. Super-sequences at each codon position across genes. (**A**) First half in translational direction (forward) and (**B**) second half opposite to translational direction (backward). Each bar, "–", represents a codon position. Vertical arrows represent super-sequences.

Figure 2. Intragenic spatial codon usage bias in prokaryotic genomes, *Escherichia coli K12* (**A** and **B**), *Bacillus subtilis* (**C** and **D**), *Thermotoga maritima* (**E** and **F**), and *Sulfolobus tokodaii* (**G** and **H**). Asterisks in the panel indicate trends are significant by isotonic regression (see Table 1). $\hat{N}_c'$ is plotted in the vertically-inverted direction because large value represents weak bias. Both the start and stop codons are excluded in the plots. For visual presentation, plots were smoothed by a locally weighted regression in a sliding window of 10 codons; however, the isotonic regression was applied to the original data. Genes were grouped into intervals with approximately equal numbers of genes by their length measured in the number of codons. The average length of each interval is given in the panel. Plot of each length interval is color-coded.

Figure 3. Trend of spatial codon usage bias as shown by the expected $\hat{N}_c'$ values from isotonic regression. Both the start and stop codons are excluded in the analysis. Side-by-side plots of expected $\hat{N}_c'$ ("+") versus codon positions for the first halves and second halves are presented along the translational direction. Lc is the gene length measured in the number of codons. Λ is the test statistic of isotonic regression. Two examples are presented, the Lc=420 interval in *E. coli* genome (**A** and **B**) and Lc=400 interval in *S. cerevisiae* genome (**C** and **D**). In *E. coli*, the

isotonic regression is applied to codon positions beyond the first 25 and before last 25 codons. In *S. cerevisiae*, regression is applied to codon positions beyond the first 20 and before the last 20 codons.

Figure 4. Incremental pattern of intragenic spatial codon usage bias in *S. cerevisiae* (**A** and **B**) and the effect of expression on the pattern (**C** and **D**). $\hat{N}_c'$ value is plotted in the inverted direction. Both the start and stop codons are excluded in the plots. Plots are smoothed by a locally weighted regression in a sliding window of 10 codons (the isotonic regression was applied to the unsmoothed data). Asterisks in the panel indicate significant trends by isotonic regression (see Table 1 and Table 2). Asterisk in parenthesis indicates a significant trend opposite to other groups. In **C** and **D**, genes are grouped into top 20%, middle 20%, and bottom 20% by their expression levels measured by Affymetrix DNA microarrays. Genes are then divided into three equal intervals by gene length (Lc). Each group is labeled by its expression level and average gene length. The bottom 20% in yeast shows flat pattern and is omitted for clarity.

Figure 5. Intragenic spatial pattern of GC content in *S. cerevisiae* and the effect of expression on the pattern. Both the start and stop codons are excluded in the plots. Plots are smoothed by a locally weighted regression in a sliding window of 10 codons. Asterisks in the panel indicate significant trends by isotonic regression (see Table 3). Genes are grouped in the same way as in Figure 5C and D.

Figure 6. The M-shaped pattern of the intragenic spatial codon usage bias in *D. melanogaster* (**A** and **B**) and the effect of expression level on the pattern (**C** and **D**). The $\hat{N}_c'$ value is plotted in the inverted direction. Asterisks in **A** and **B** indicate significant trends in the middle section by isotonic regression (see Table 1). Plots are smoothed by a locally weighted regression in a sliding window of 10 codons (the isotonic regression was applied to the unsmoothed data.). Genes are grouped by gene length. Each group contains similar numbers of genes and is labeled by its average gene length (Lc). Plot of each length interval is color-coded.

Figure 7. The intragenic spatial pattern of GC content variation in *D. melanogaster.* GC content is estimated using the first and second base of each codon position. Genes are first partitioned by expression level as in Figure 6C and D and then grouped by gene length measured in the number of codons (Lc).

Figure 8. The effect of introns on the intragenic spatial patterns of codon usage bias (**A** and **B**) and GC content variation (**C** and **D**) in *D. melanogaster*. Only the longest groups of genes are presented. Equal numbers of intronless and two-exoned genes are used to ensure appropriate comparisons.

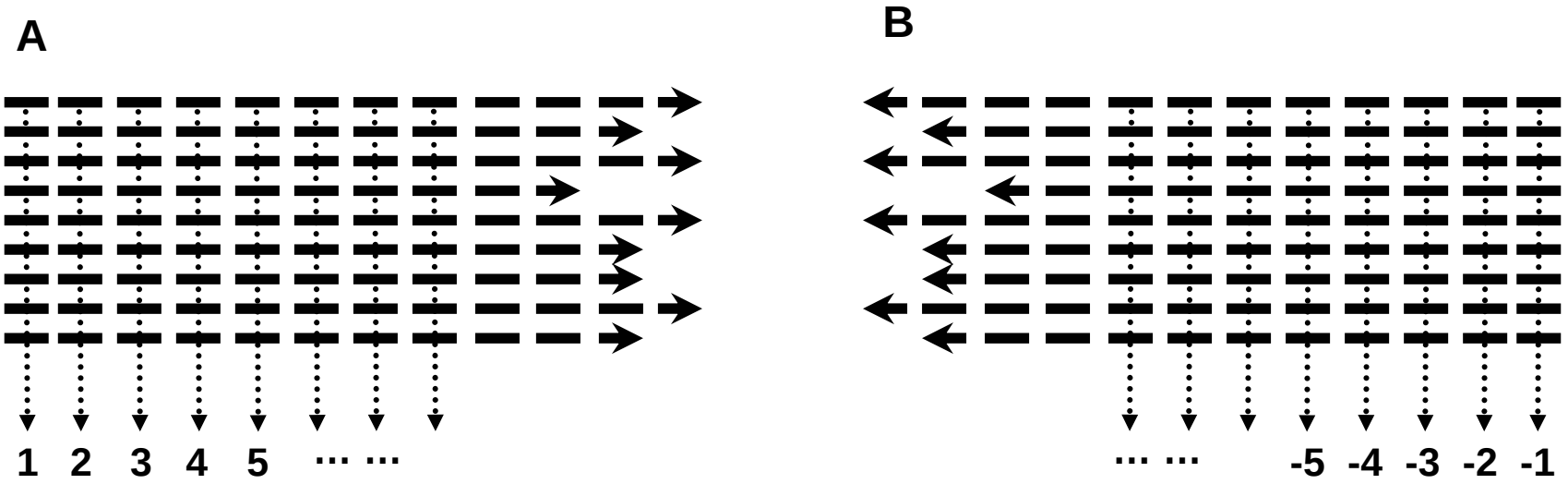


Figure 1A and B

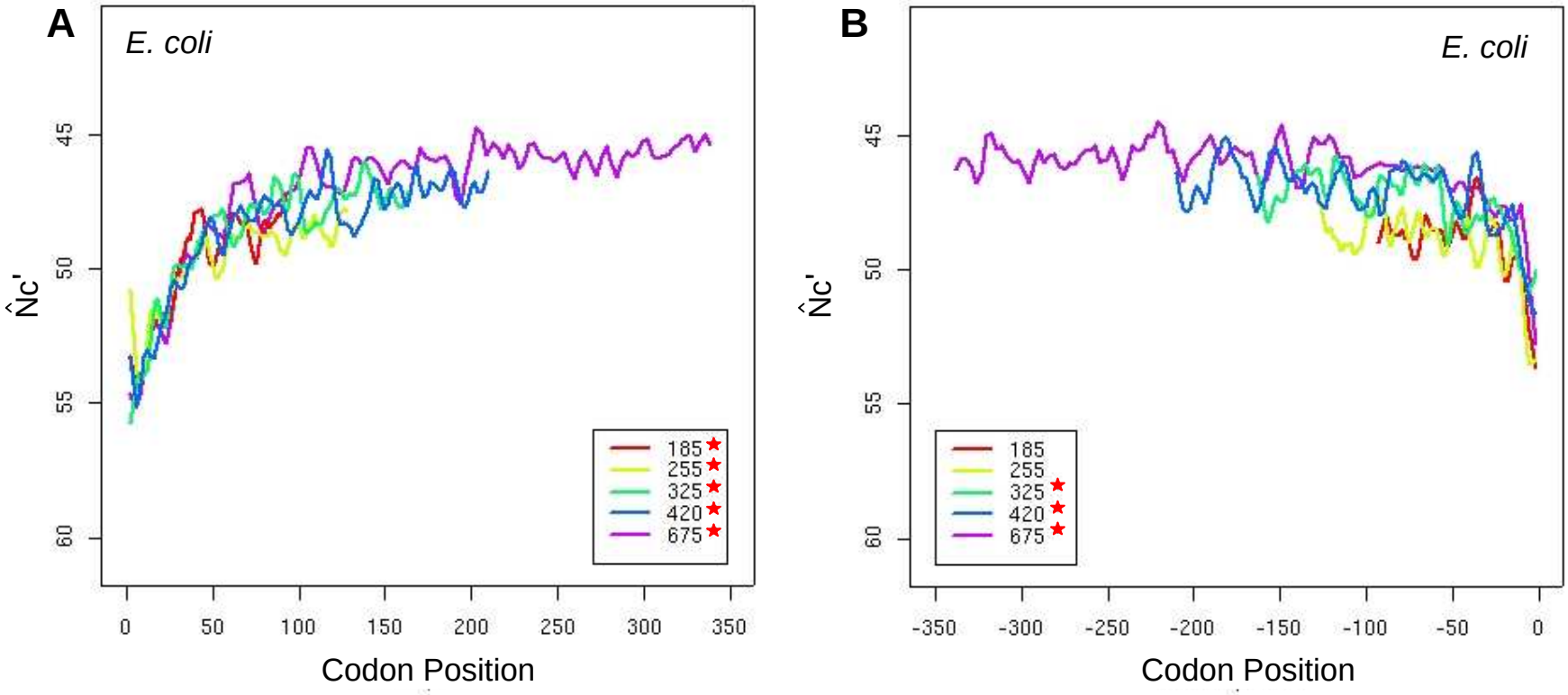


Figure 2A and B

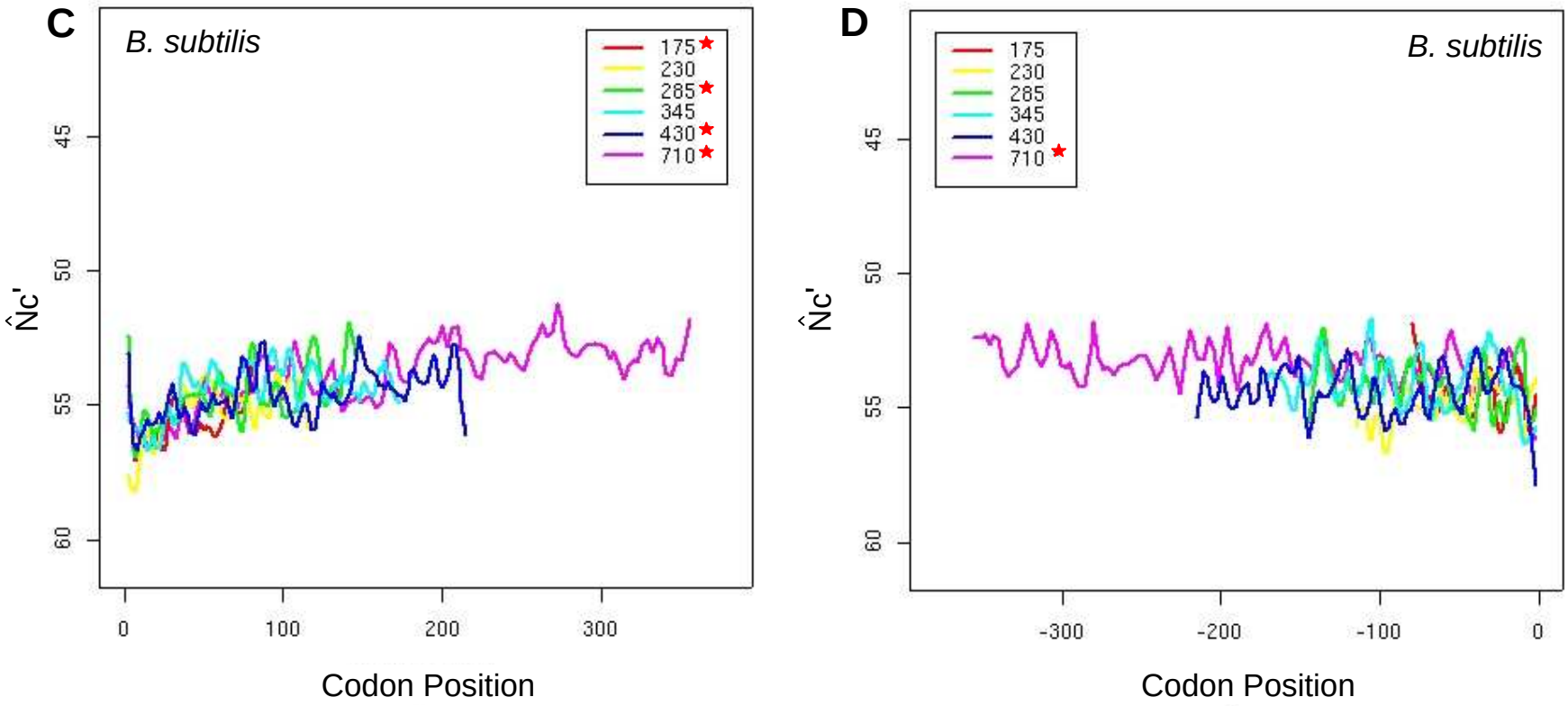


Figure 2C and D

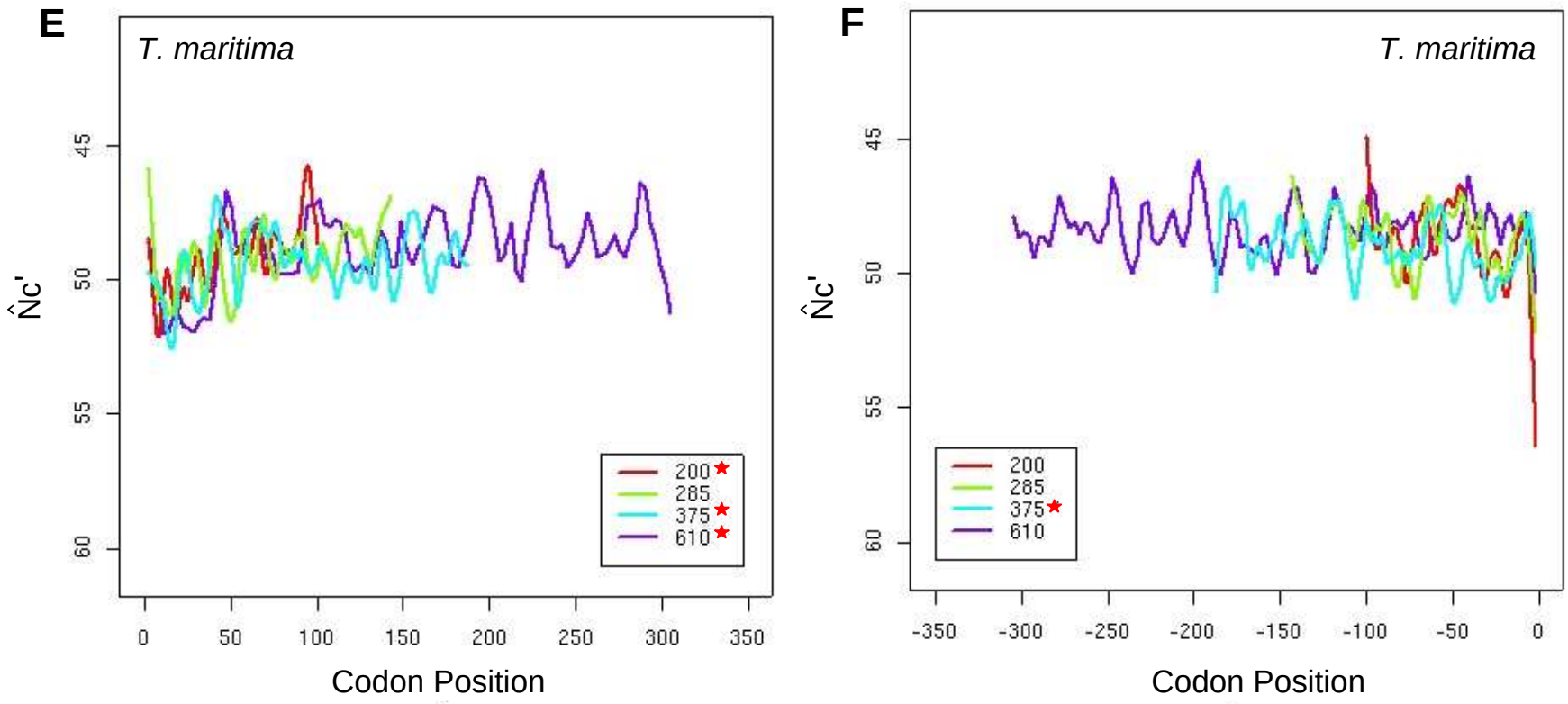


Figure 2E and F

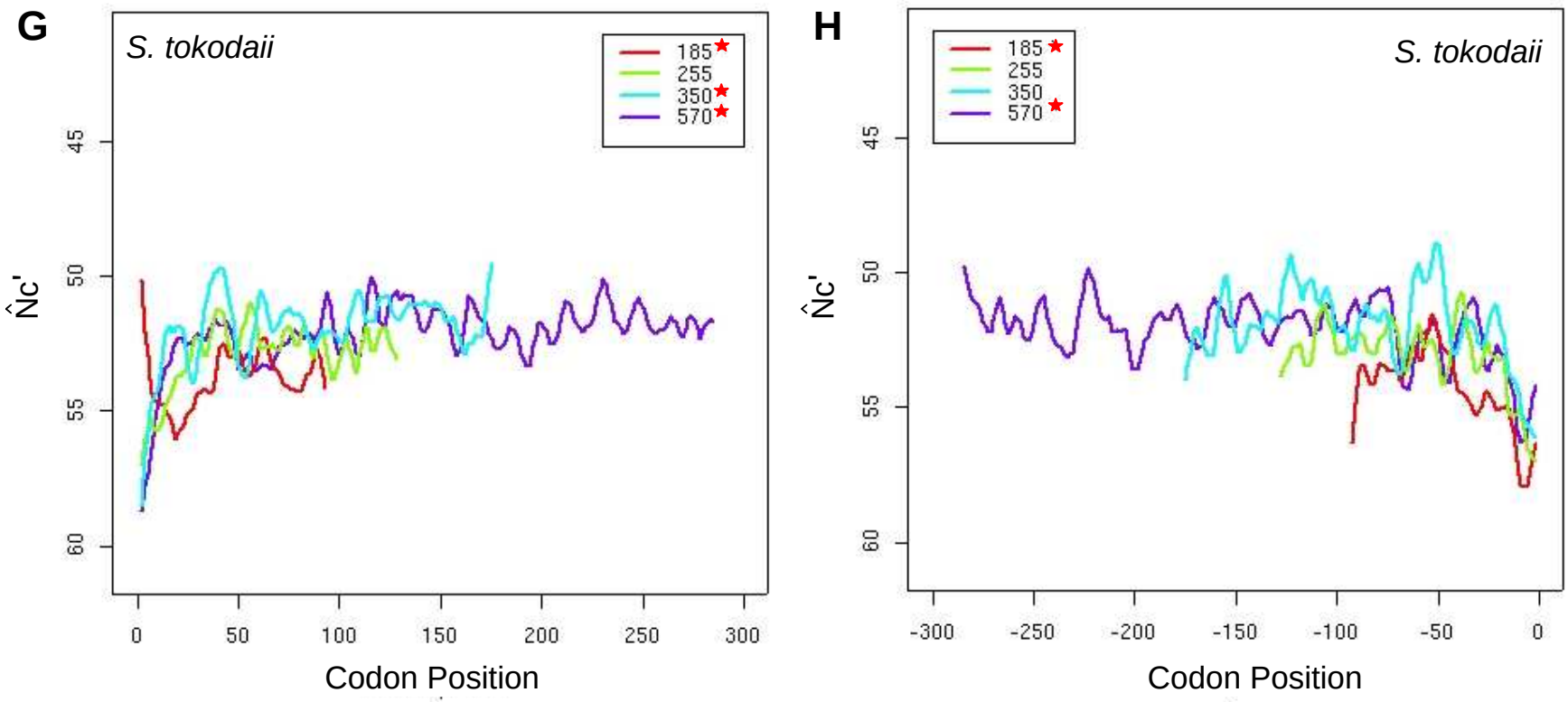


Figure 2G and H

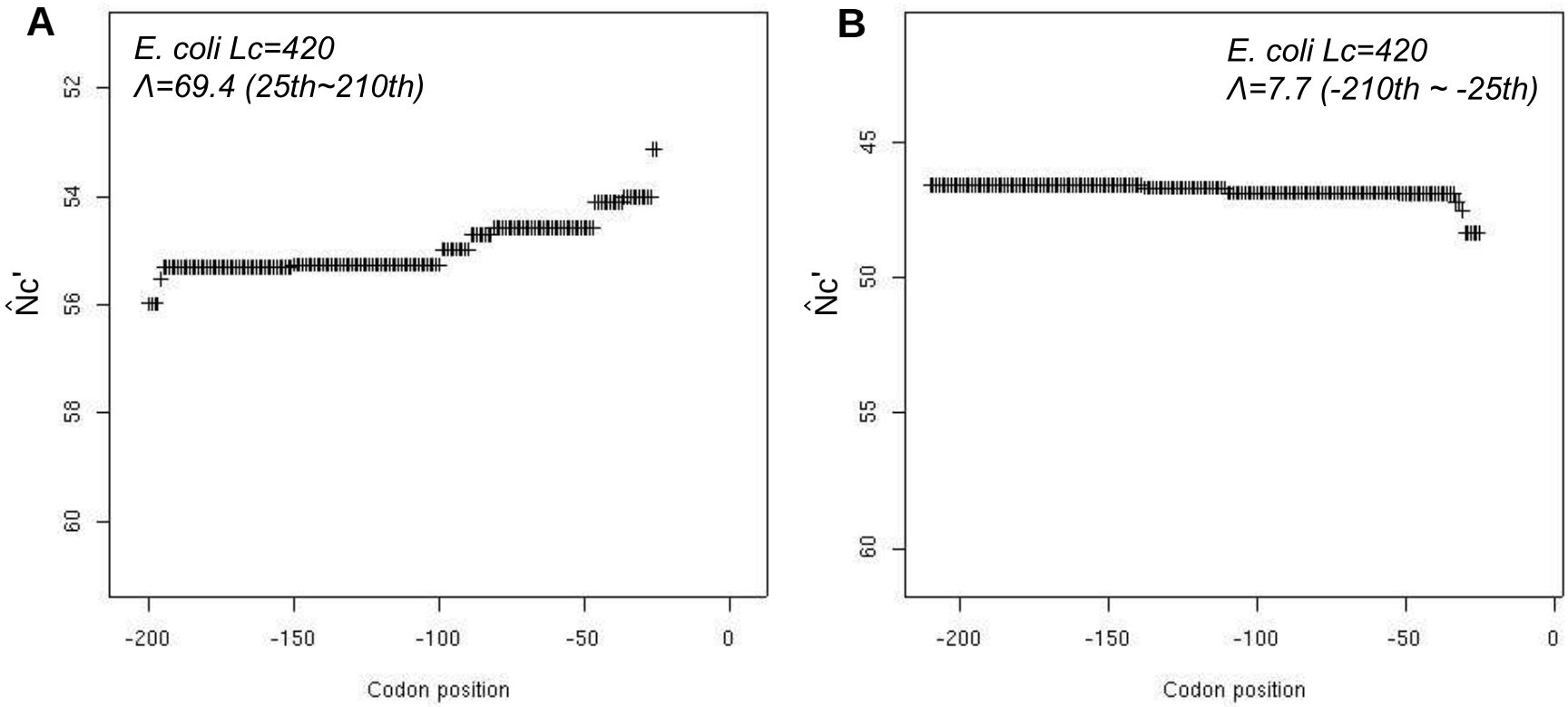


Figure 3A and B

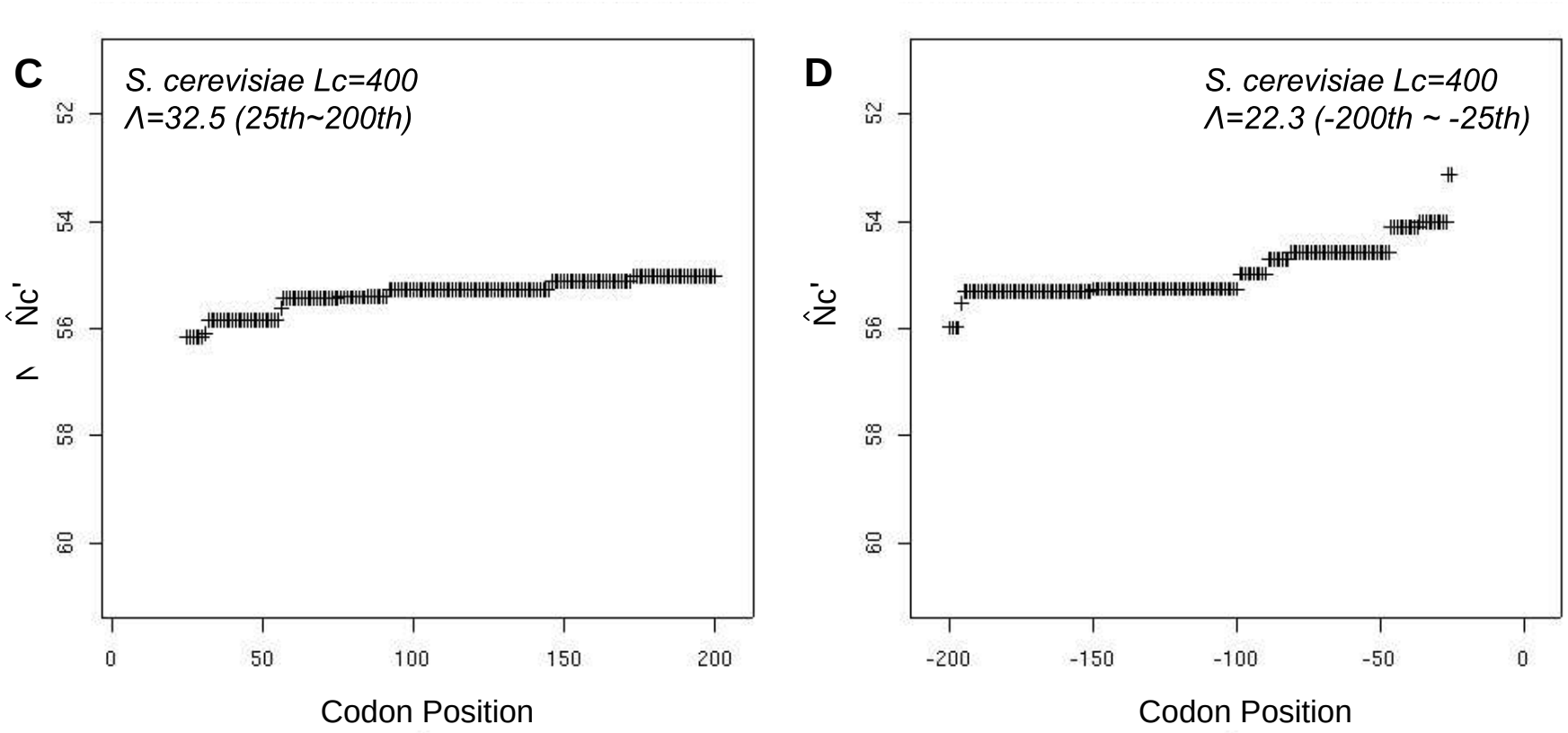


Figure 3C and D

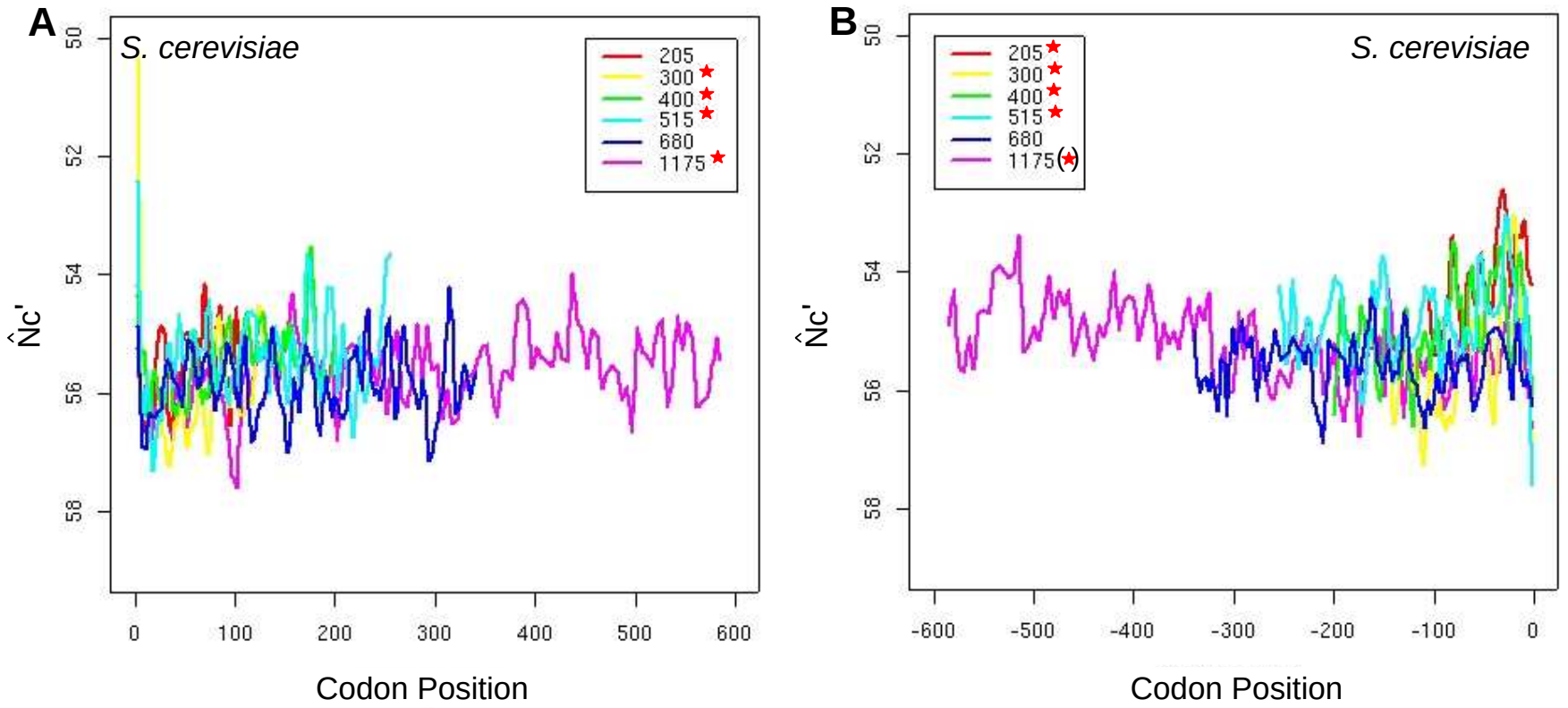


Figure 4A and B

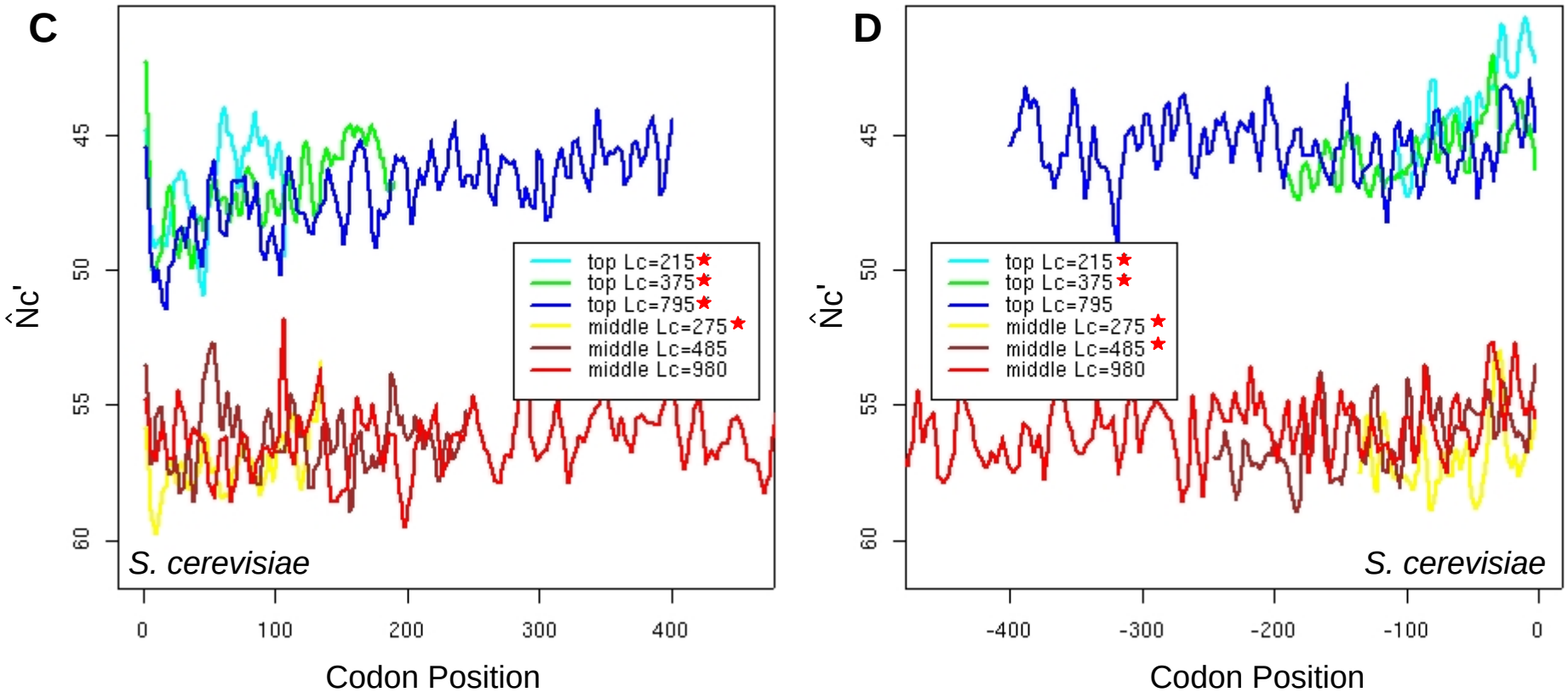


Figure 4C and D

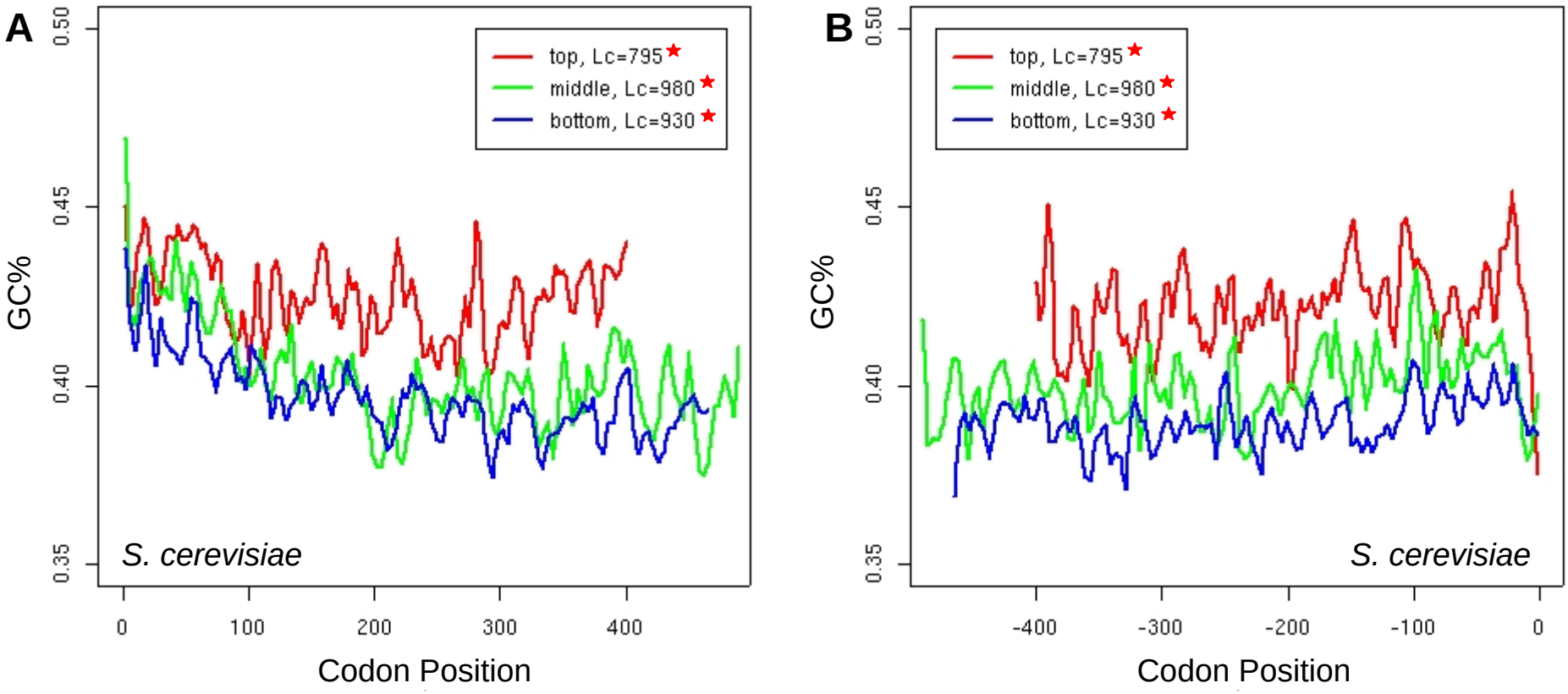


Figure 5A and B

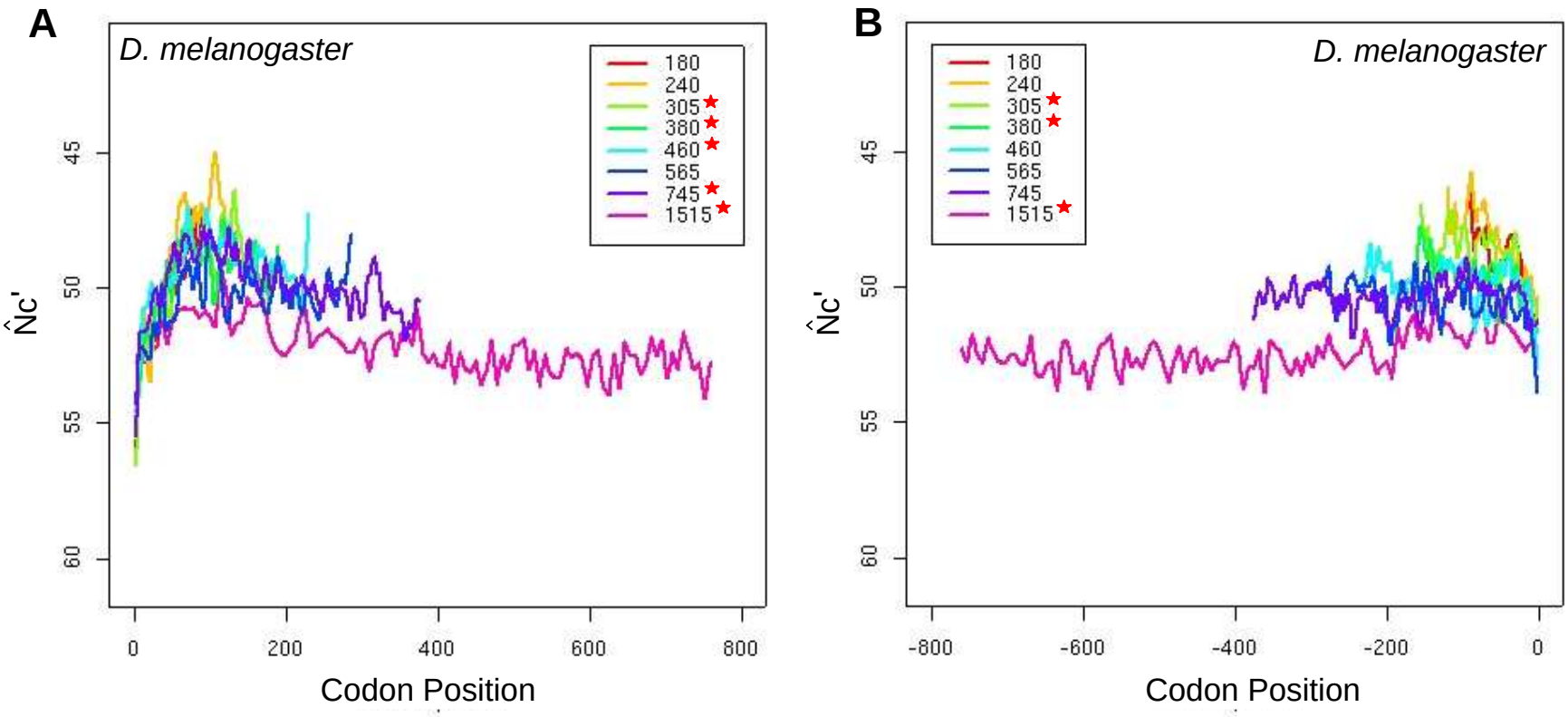


Figure 6A and B

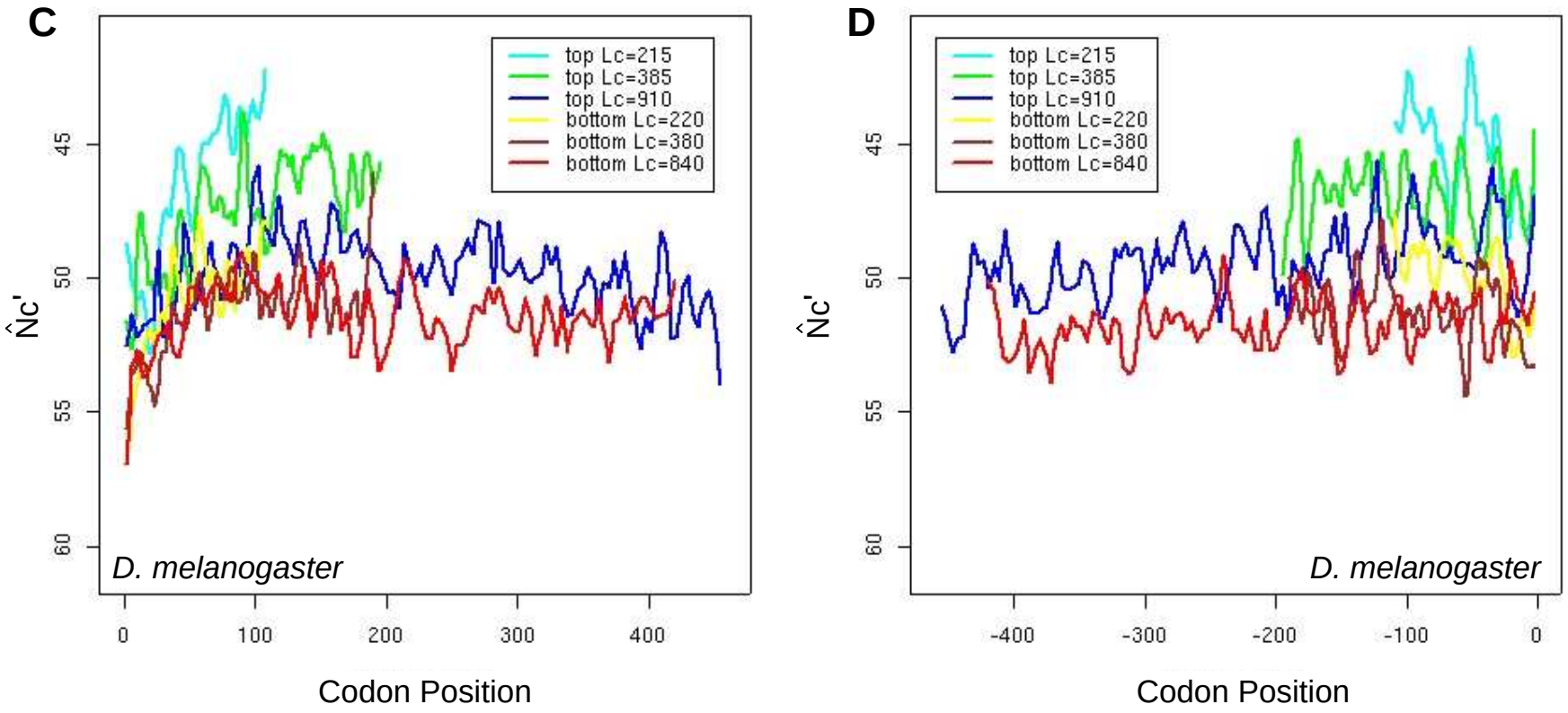


Figure 6C and D

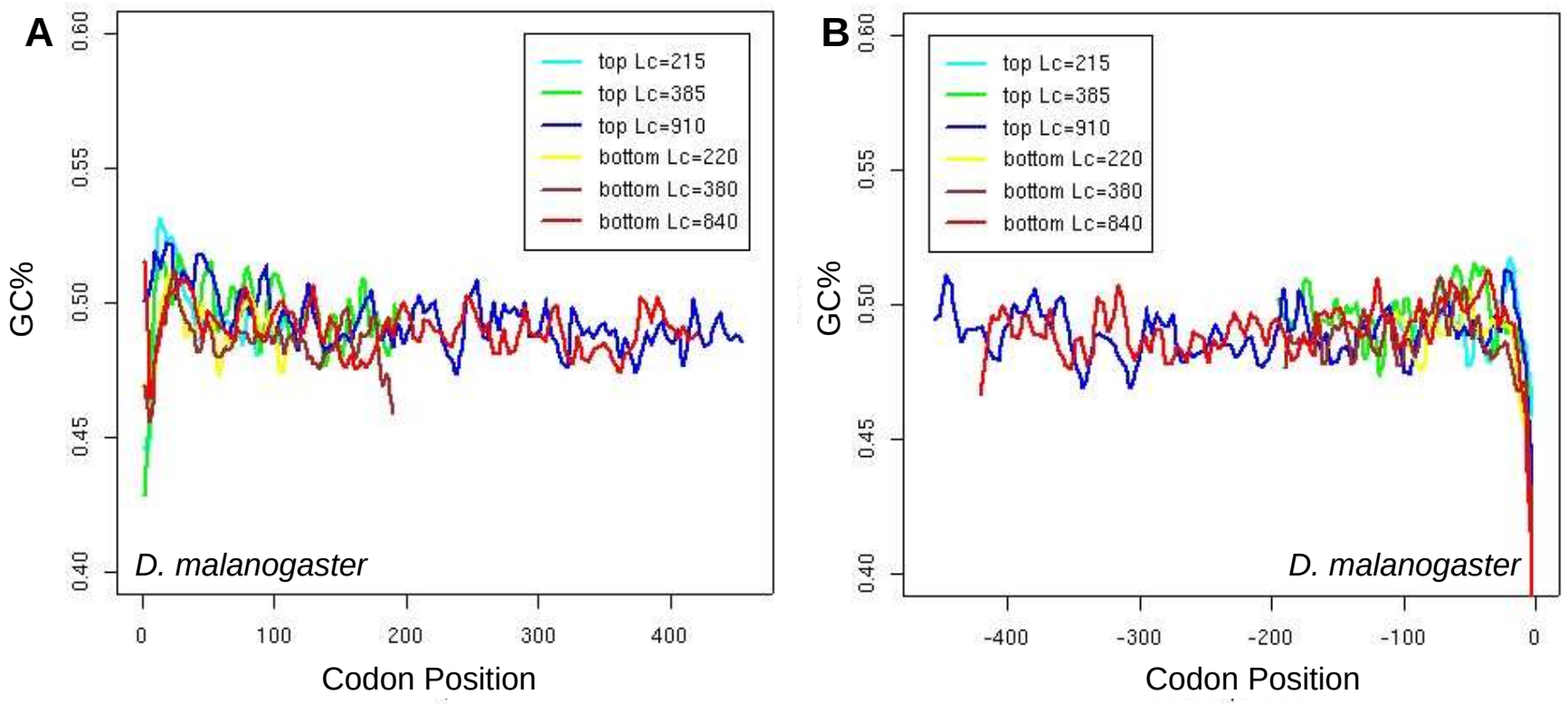
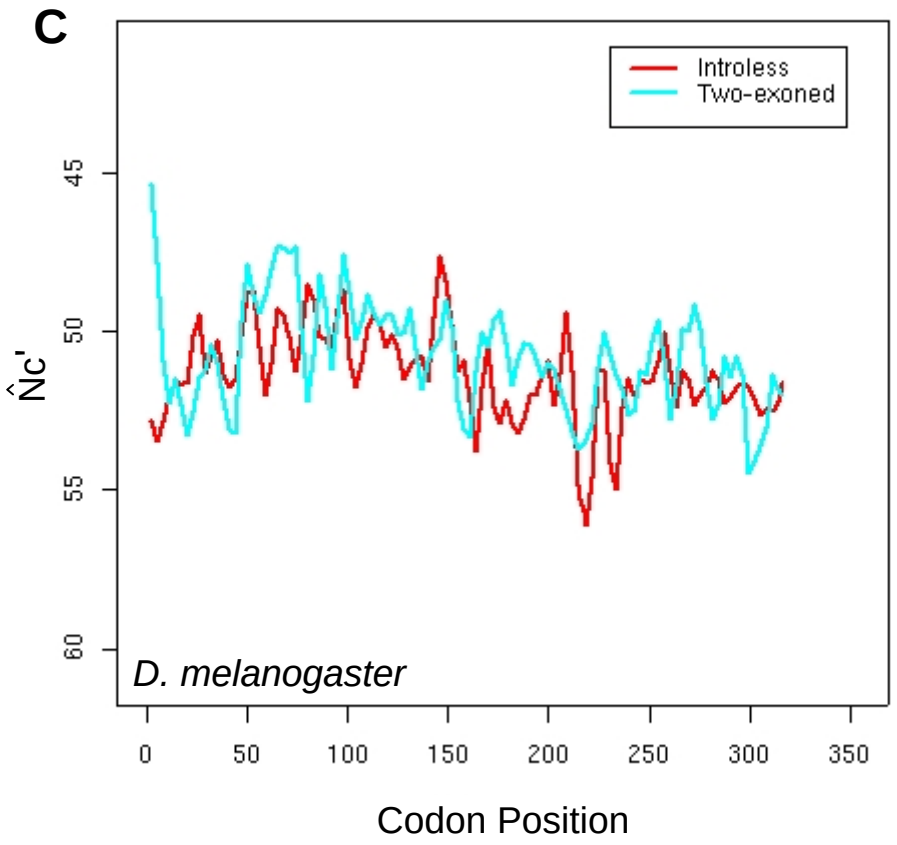


Figure 7A and B



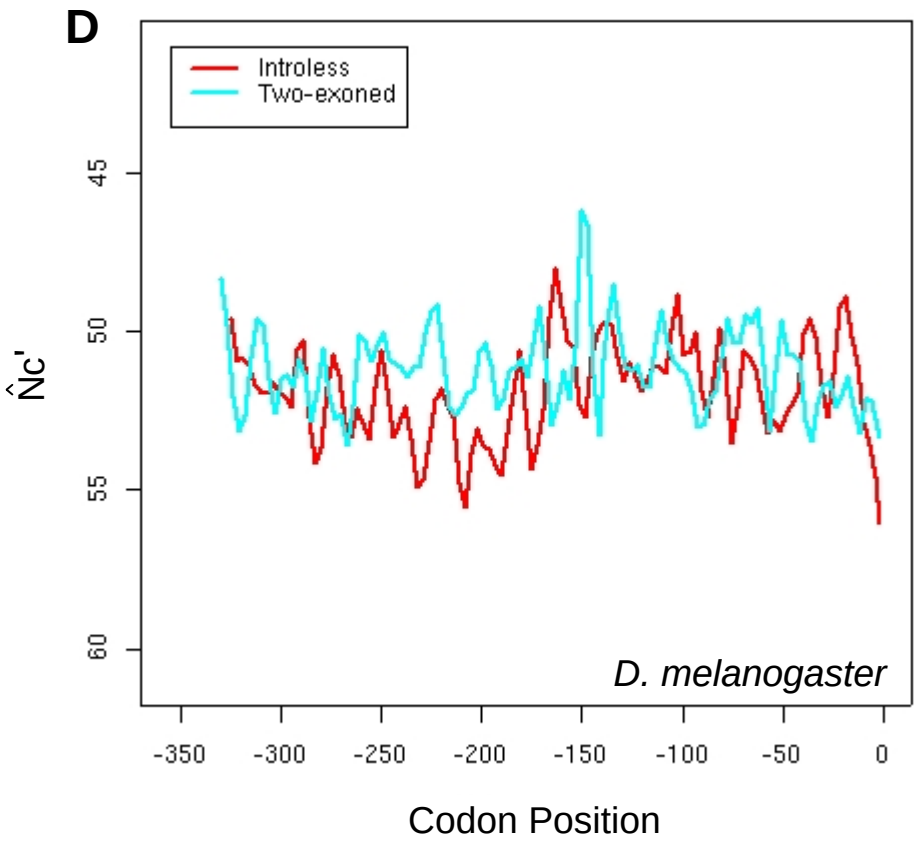


Figure 8A and B

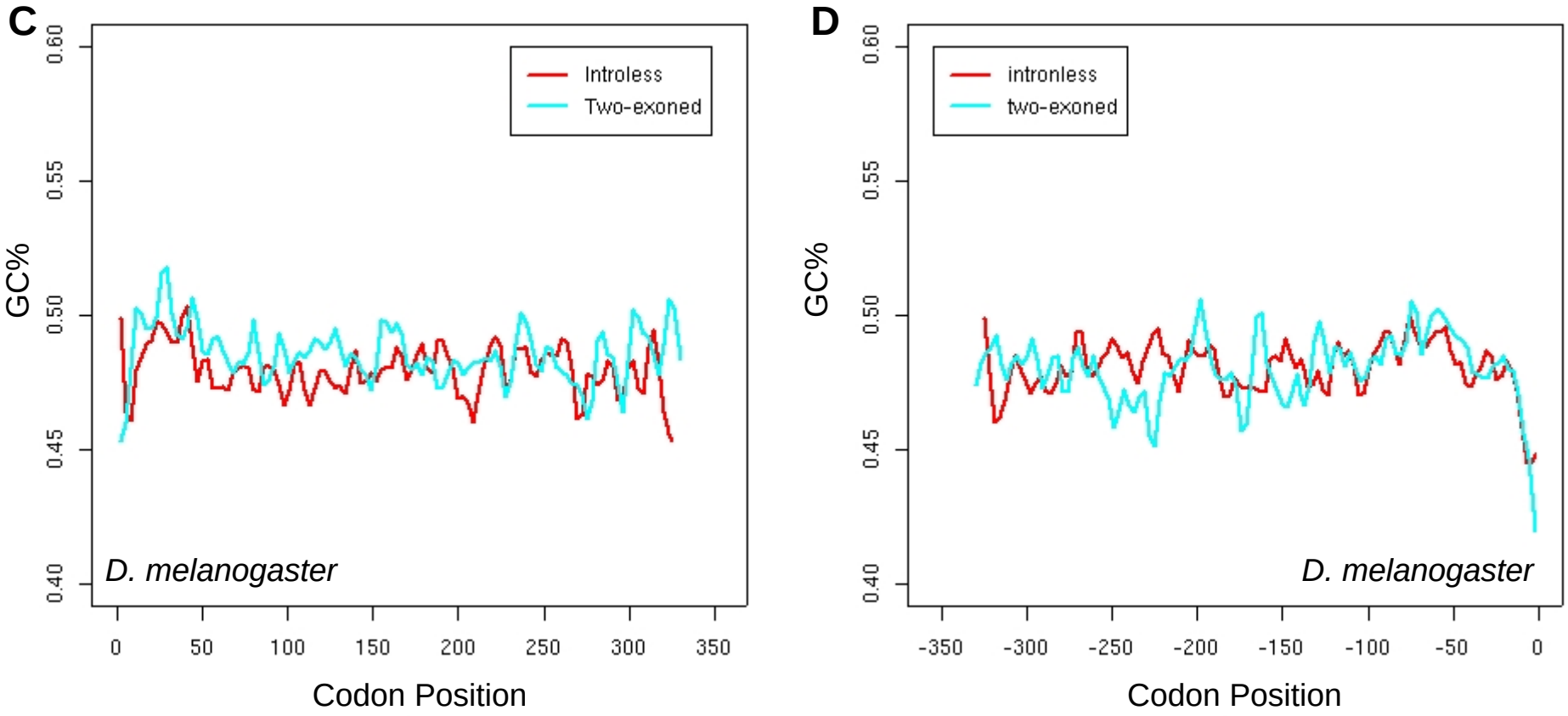


Figure 8C and D

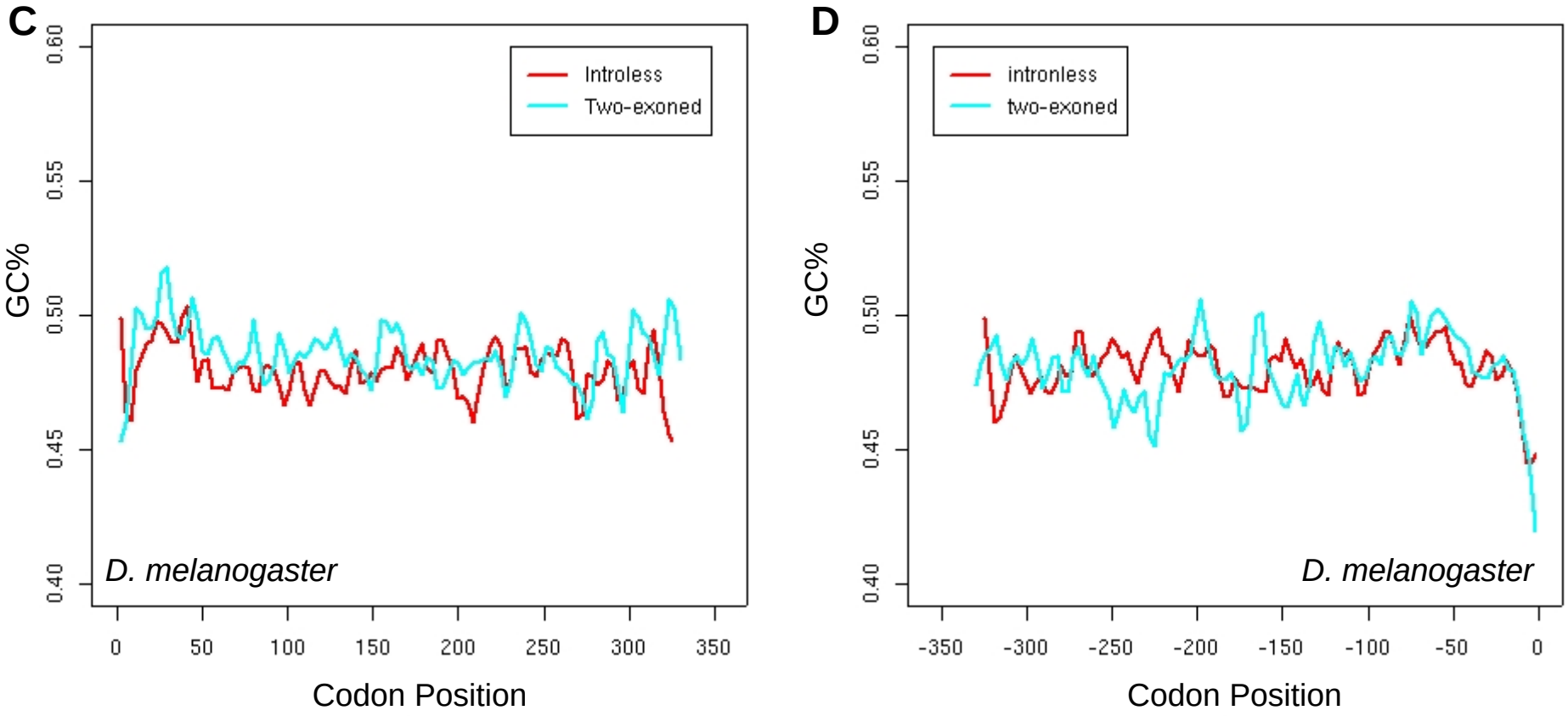


Figure 8C and D