Genome-Wide Selection on Codon Usage at the Population Level in the Fungal Model Organism Neurospora crassa

C.A. Whittle, Y. Sun, and H. Johannesson*

Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

*Corresponding author: E-mail: hanna.johannesson@ebc.uu.se.

Abstract

Many organisms exhibit biased codon usage in their genome, including the fungal model organism Neurospora crassa. The preferential use of subset of synonymous codons (optimal codons) at the macroevolutionary level is believed to result from a history of selection to promote translational efficiency. At present, few data are available about selection on optimal codons at the microevolutionary scale, that is, at the population level. Herein, we conducted a large-scale assessment of codon mutations at biallelic sites, spanning more than 5,100 genes, in 2 distinct populations of N. crassa: the Caribbean and Louisiana populations. Based on analysis of the frequency spectra of synonymous codon mutations at biallelic sites, we found that derived (nonancestral) optimal codon mutations segregate at a higher frequency than derived nonoptimal codon mutations in each population; this is consistent with natural selection favoring optimal codons. We also report that optimal codon variants were less frequent in longer genes and that the fixation of optimal codons was reduced in rapidly evolving long genes/proteins, trends suggestive of genetic hitchhiking (Hill-Robertson) altering codon usage variation. Notably, nonsynonymous codon mutations segregated at a lower frequency than synonymous nonoptimal codon mutations (which impair translational efficiency) in each N. crassa population, suggesting that changes in protein composition are more detrimental to fitness than mutations altering translation. Overall, the present data demonstrate that selection, and partly genetic interference, shapes codon variation across the genome in N. crassa populations.

Key words: Neurospora crassa, microevolution, codon, selection, genetic hitchhiking.

Introduction

Neurospora crassa comprises a major model system for genetics and evolutionary research. Similar to many organisms (e.g., Caenorhabditis, Populus, Silene, Drosophila, and others, Duret and Mouchiroud 1999; Duret 2000; Maside et al. 2004; Cutter et al. 2008; Ingvarsson 2008; Qiu et al. 2011a), species of Neurospora exhibit biased codon usage at the macroevolutionary scale. For example, elevated levels of a specific subset of codons have been found in the genome of N. crassa (Whittle et al. 2011a). In addition, the usage of these codons has been associated with elevated gene expression in its close relatives N. discreta and N. tetrasperma (Whittle et al. 2011a, 2011b). Thus, the available species-level data are consistent with the theory that selection has historically favored the usage of a specific set of optimal codons that confer a selective advantage for efficient and accurate translation in Neurospora (Ikemura 1981; Gouy and Gautier 1982).

Although biased codon usage has been reported at the macroevolutionary scale in many species, few data are available about selection for optimal codon usage at the microevolutionary scale, that is, at the population level (Akashi and Schaeffer 1997; Maside et al. 2004). One method used to assess selective differences among classes of mutations, such as optimal and nonoptimal codons, is to compare their frequency distributions in a population (Sawyer et al. 1987; Akashi and Schaeffer 1997). The advantage of this method is that at the population level, each class of mutations should be similarly affected by population structure, demography, and evolutionary history (Hudson 1993; Akashi and Schaeffer 1997; Akashi 1999); thus differences in their frequency distributions can be attributed to fitness effects. According to the mutation–selection–drift model (Kimura 1983), the distribution of derived optimal codon mutations (i.e., mutations resulting in changes from nonoptimal to optimal codons) should skew more toward high-frequency variants than derived nonoptimal codon mutations, due to their selective advantage (Akashi and Schaeffer 1997; Maside et al. 2004). Based on this principle, analyses of polymorphic sites in Drosophila and plants, including two genes of Drosophila pseudoobscura, 14 genes of D. mauritiana, between 8 and 14 genes of D. simulans, 18 genes of D. americana, and 18 genes in Silene latifolia, have revealed ongoing selection favoring optimal codons (Akashi and Schaeffer 1997; Maside et al. 2004; Llopart et al. 2008; Qiu et al. 2011a). The findings affirm variation in fitness classes of synonymous mutations within populations. At present, however, such assessments at the population level have not been conducted at the genome-wide level or for other major taxonomic groups, including the key fungal model system Neurospora.

Neurospora crassa provides a particularly useful model system to study selection on codon usage in populations. As in most filamentous ascomycetes, each individual of N. crassa contains haploid nuclei with a single mating type within its mycelium. During the sexual cycle, fertilization results in a short diploid stage, immediately followed by...
meiosis (Shear and Dodge 1927). Thus, N. crassa is likely to be exposed to extensive haploid selection, a feature that may enhance the efficiency of removal of deleterious mutations and the fixation of advantageous genomic changes (particularly in large populations), which is in contrast to diploid or polyploid systems (Greg and Travisano 2003; Zeyl et al. 2003). The lack of sheltering and the exposure of phenotypes to selection expected for N. crassa make it an attractive system to study the distribution of mutational classes in populations. The study of genome evolution in N. crassa is facilitated by the fact that its genome has been fully sequenced and well-characterized; the genome contains approximately 10,000 genes and 43 Mbp DNA across 7 chromosomes (Galagan et al. 2003). In addition, a large-scale transcriptome data set of N. crassa populations has recently become available (Ellison et al. 2011), making investigations of codon usage possible at the population level.

One key genomic feature that should be assessed in conjunction with an evaluation of codon usage within populations is gene length. For example, it has been reported that shorter genes have greater biases in codon usage than longer genes in certain animals and plants, for example, Drosophila, Arabidopsis, Silene, and species of Caenorhabditis (Duret and Mouchiroud 1999; Cutter et al. 2008; Qui et al. 2011a). An inverse association between gene length and codon usage bias has also recently been reported for species of Neurospora, including N. tetrospерma and N. discreta (Whittle et al. 2011b). The gene-length findings might result from Hill–Robertson effects in longer genes (Hill and Robertson 1966), wherein genetic hitchhiking interferes with the selection on genetically linked mutations (Comeron and Guthrie 2005; Loewe and Charlesworth 2007). For instance, recurrent positive selection events may drag linked weakly selected synonymous codon mutations to fixation (Rice 1987) whereas negative selection may force linked beneficial mutations to be eliminated (background selection, Charlesworth et al. 1993); this may lower population genetic diversity and hinder the fixation of optimal codons (Charlesworth et al. 1993; Loewe and Charlesworth 2007). In this regard, an evaluation of codon usage in N. crassa populations requires consideration of the possible role of gene length.

In the present investigation, we tested for ongoing selection on codon usage at the population level in N. crassa using the large-scale biallelic single nucleotide polymorphism (SNP) data set presented by Ellison et al. (2011). Based on these data, Ellison et al. (2011) revealed two well-defined N. crassa populations: the tropical Caribbean population and the subtropical Louisiana population (N = 20). In our analysis, we identified the codons located at every SNP position in each of these populations, assessed the number of ancestral and derived codons at each site using two Neurospora species as outgroups, and compared the frequency distributions of derived optimal and derived nonoptimal codons. In addition, we evaluated the frequency distribution of nonsynonymous codon mutations as compared with synonymous mutations and assessed the role of gene length in the evolution of codon usage in each N. crassa population.

Materials and Methods

The N. crassa Caribbean and Louisiana populations have been previously defined and characterized by Ellison et al. (2011) using their complete transcriptome SNP data set. Based on their analyses, it has been estimated that the populations diverged approximately 0.4 Ma, have about 9.4% fixed nucleotide differences, and likely diverged without complete allopatry as they exist within a 1,000 km proximity in some regions. Furthermore, nucleotide diversity, the average number of pairwise differences among individuals ([rt]), has been calculated as 0.0023 (Caribbean) and 0.0024 (Louisiana). Approximately 34% of nonsynonymous sites have been estimated to be deleterious. The Louisiana population is endemic to this region whereas the Caribbean population spans Florida, Haiti, and the Yucatan (Ellison et al. 2011).

Generation of the Biallelic Codon Data Set

The codons located at biallelic sites in the Caribbean and Caribbean N. crassa populations were determined using the publicly available SNP data file that spans 133,108 biallelic SNP positions (Ellison et al. 2011; the version available at http://pmb.berkeley.edu/~taylor/ftp/Ellison_2011_SNPdata.txt downloaded May 2011). The data set contains two pieces of information: the genomic location of the SNPs in the reference N. crassa genome (N. crassa OR74A [NC10]) and the nucleotide found at these positions in each individual. The numbers of individuals sampled were 19 and 20 for the Caribbean and Louisiana populations, respectively. Using these data, we identified all SNPs that were located within the protein-coding (coding sequences [CDS]) region of genes in N. crassa. Subsequently, we determined the codons associated with each SNP for every individual. The variable sites are referred to as biallelic codon sites herein. Each N. crassa biallelic codon site was classified as synonymous or nonsynonymous. Synonymous biallelic codon sites were those that contained two synonymous codons encoding a single amino acid. Nonsynonymous sites contained codons encoding different amino acids.

Identification of Outgroup Sequences

Two close relatives of N. crassa, N. tetrospерma, and N. discreta, from which genomic data have recently become available (http://www.jgi.doe.gov) were used as outgroups. Previous phylogenetic analyses have shown that N. tetrospерma (NT) and N. crassa (NC) form a clade relative to N. discreta (ND), that is, the phylogenetic relationship is (INT, NC), ND (Dettman et al. 2003). The pairwise divergences (dS and dN values) have been estimated respectively as 0.055 and 0.004 for NC and NT, 0.156 and 0.015 for NC and ND, and 0.1481 and 0.015 for NT and ND (from concatenated genes with complete CDS identified herein and using the LPB93 method in PAML, Yang 2007). The genome sequences for these outgroup taxa were used to identify homologous codons for each biallelic codon site in N. crassa. In order to identify the homologous
Table 1. The Number of Synonymous and Nonsynonymous Biallelic Codon Sites in the Caribbean and Louisiana Neurospora crassa Populations and the Number of Positions with an Ancestral Codon Identified in N. tetrasperma and N. discreta.

<table>
<thead>
<tr>
<th>Status of SNP</th>
<th>Total No. of Biallelic Codon Sites in N. crassa</th>
<th>Sites with Homologous Codon Identified in N. tetrasperma and N. discreta</th>
<th>Sites with a Defined Ancestral Codon*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Caribbean</td>
<td>Louisiana</td>
<td>Caribbean</td>
</tr>
<tr>
<td>Biallelic</td>
<td>57,564b</td>
<td>61,292b</td>
<td>51,404</td>
</tr>
<tr>
<td>Synonymous sites</td>
<td>46,764</td>
<td>50,235</td>
<td>42,404</td>
</tr>
<tr>
<td>Nonsynonymous sites</td>
<td>10,800</td>
<td>11,057</td>
<td>9,000</td>
</tr>
<tr>
<td>Number of genes with biallelic codon sites</td>
<td>5,244</td>
<td>5,182</td>
<td>4,771</td>
</tr>
</tbody>
</table>

* A position wherein identical codons were found in N. discreta and N. tetrasperma.

b The number of SNP sites that are uniallelic within the Caribbean and within the Louisiana populations are 25,198 and 21,470. These values added to biallelic sites using the outgroup taxa.

codon in N. tetrasperma and N. discreta corresponding to each biallelic codon site in N. crassa, the complete genomic sequences for all three taxa were aligned and SNPs and associated codons were extracted using the SNP position data file from Ellison et al. (2011) and the Blat program (Kent 2002). The full gene CDS sequence containing one or more biallelic sites were isolated for each of the three taxa; the three CDS sequences per gene were aligned and gaps removed using Blat and were used for CDS-level comparisons.

Designations of Codon Status at Neurospora crassa Biallelic Sites

We designated the ancestral codons for the N. crassa biallelic codon sites using the outgroup taxa. In particular, ancestral codons were defined for sites wherein the codons were identical in both N. tetrasperma and N. discreta. Sites wherein an ancestral codon was not defined were excluded from further analysis. This approach is more conservative than using a single outgroup taxon and greatly limits the misidentification of ancestral codons, which can impede accurate detection of selective effects in the population (Akashi and Schaeffer 1997). Derived codons were defined as those that differed from the ancestral codon. Using this approach, the two codons located at each N. crassa biallelic codon site were categorized as an ancestral codon or a derived codon.

For each synonymous biallelic codon site in N. crassa, we classified the codons as optimal or nonoptimal. The optimal codon list for N. crassa was assumed to be equivalent to that reported for N. tetrasperma and N. discreta, which represent the codons most frequently used in highly expressed genes (supplementary table S1, Supplementary Material online; Whittle et al. 2011b). This codon list is highly similar to the most frequent (preferred) codons within the N. crassa genome (Whittle et al. 2011a). Using this codon list, the optimal and nonoptimal codons were designated for each synonymous biallelic site in N. crassa and for the homologous ancestral codon identified from the outgroup taxa.

We assessed the frequency distributions of derived optimal versus derived nonoptimal codon mutations at synonymous biallelic sites in each N. crassa population. For this, derived optimal codon mutations were defined as those involving a switch from an ancestral nonoptimal codon to an optimal codon at an N. crassa biallelic site. Derived nonoptimal codon mutations involved a switch of an ancestral optimal codon to a nonoptimal codon at a biallelic site. For nonsynonymous biallelic codon sites, derived mutations were identified as those mutations involving a change in an amino acid (relative to the ancestral state), regardless of the optimal codon status.

Molecular Evolutionary Analyses

For our population-level molecular evolutionary analyses, we compared the frequency distribution of derived optimal and derived nonoptimal codons across all biallelic codon sites within each N. crassa population, in order to test for selection on codon usage. In conjunction with this, we evaluated the role of gene length and compared the frequency profiles of nonsynonymous changes as compared with synonymous changes in order to assess fitness differences among these classes of mutations. Complementary macroevolutionary scale analyses in N. crassa (relative to outgroup taxa) were conducted on protein evolution relative to the frequency of optimal codons (Fop) and of GC content of introns (GCI) versus exons in order to further reveal factors driving codon usage in this taxon. All statistical analyses were conducted using SigmaStat 3.5 (http://www.sigmaplot.com).

Results

We converted the N. crassa data set of 133,108 biallelic SNPs by Ellison et al. (2011) into a codon polymorphism data set for two distinct populations: the Caribbean (N = 19) and Louisiana populations (N = 20). After isolation of all SNPs that were located within the protein-coding (CDS) regions of genes in N. crassa, we found a total of 82,762 SNPs that were located within 5,418 genes across the 2 populations.

A total of 51,404 and 54,595 of the biallelic codon sites in the N. crassa Caribbean and Louisiana data sets, respectively, had a homologous codon identified in both N. discreta and N. tetrasperma (table 1). Based on the criteria for defining ancestral codons outlined above, we found a total of 37,934 biallelic codon sites in the N. crassa Caribbean data set and 40,421 sites from the Louisiana data set that had a defined ancestral codon (table 1). A total of 31,710
and 34,082 of the biallelic codon sites (with defined ancestral codons) contained synonymous codons whereas 6,224 and 6,339 contained nonsynonymous codons in the Caribbean and Louisiana data sets, respectively (table 1).

For each N. crassa biallelic codon site (with a defined ancestral state), we identified the ancestral and derived codons for each population. The data reveal that the nearly half of the codons located within synonymous biallelic sites were ancestral codons. For instance, among the 63,420 codons identified within the 31,710 synonymous biallelic codon sites from the Caribbean data set, a total of 31,021 were ancestral codons (48.9%) whereas 32,399 (51.1%) were derived codons (table 2). Similar findings were obtained for the Louisiana data set, wherein 48.9% were ancestral codons and 51.1% were derived codons (table 2). The fraction of synonymous biallelic codon sites that contained one ancestral codon and one derived codon was high, representing 97.8% (31,021/31,710) of sites in the Caribbean data set and 97.9% (33,370/34,082) of sites in the Louisiana data set (table 2) whereas the remaining <2.2% of sites from each population contained two derived codons (table 2). Thus, it is evident that a single switch from an ancestral to derived codon is responsible for most of these polymorphisms in the N. crassa populations. For nonsynonymous codon sites, the ratio of ancestral to derived codons was marginally lower than for synonymous sites, for example, 46.4% of codons located at sites in the Caribbean data set were ancestral codons whereas 53.6% were derived codons (table 2), possibly reflecting selective differences among synonymous and nonsynonymous sites.

Derived Optimal and Nonoptimal Codons
A primary focus of our analyses was to assess derived optimal and derived nonoptimal codon mutations at biallelic sites in each N. crassa population, as these are the most informative about selective differences at synonymous sites (Akashi and Schaeffer 1997; Akashi 1999). Among all derived codons identified at synonymous biallelic sites in N. crassa, slightly more than half (between 51.2% and 51.6%) were derived from a mutation resulting in a switch between optimal (OP) and nonoptimal codons (NOP), for example, OP to NOP or NOP to OP, whereas the remainder involved transitions among optimal codons or among nonoptimal codons (table 2). The latter classes of mutations were excluded from further analyses as there is no prediction about their fitness effects. We found that the derived optimal codon mutations (NOP to OP) were less commonly represented than derived nonoptimal codon mutations (OP to NOP) among the two alternate codons located at biallelic sites. For example, the total number of derived optimal codons was almost half that of the number of derived nonoptimal codons located within biallelic codon sites (table 2). Nevertheless, examination of the frequency distribution of derived optimal and derived nonoptimal codons within the N. crassa populations is key to revealing whether selective differences exist among these two categories of mutations.

### Table 2. The Number of Ancestral and Derived Codons among Biallelic Codon Sites in the Caribbean and Louisiana Populations of *Neurospora crassa*.

<table>
<thead>
<tr>
<th>Categories for N. crassa Data Set</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Caribbean</td>
</tr>
<tr>
<td>Number of synonymous biallelic codon sites</td>
<td>31,710</td>
</tr>
<tr>
<td>Ancestral codons</td>
<td>31,021</td>
</tr>
<tr>
<td>Derived codons</td>
<td>32,399</td>
</tr>
<tr>
<td>Derived codons: NOP to OP&lt;sup&gt;a&lt;/sup&gt;</td>
<td>5,399</td>
</tr>
<tr>
<td>Derived codons: OP to NOP&lt;sup&gt;b&lt;/sup&gt;</td>
<td>11,341</td>
</tr>
<tr>
<td>Other derived categories&lt;sup&gt;d&lt;/sup&gt;</td>
<td>15,659</td>
</tr>
<tr>
<td>Number of nonsynonymous biallelic codon sites</td>
<td>6,224</td>
</tr>
<tr>
<td>Ancestral codons</td>
<td>5,772</td>
</tr>
<tr>
<td>Derived codons&lt;sup&gt;c&lt;/sup&gt;</td>
<td>6,676</td>
</tr>
<tr>
<td>Total number of biallelic codon sites</td>
<td>37,934</td>
</tr>
</tbody>
</table>

**Note:** Ancestral codons were defined at those positions wherein identical codons were found in *Neurospora discreta* and *Neurospora tetrasperma*.

<sup>a</sup> The number of derived codons at synonymous biallelic sites that have shifted from NOP in the ancestral codon to OP at the biallelic site.

<sup>b</sup> The number of derived codons at synonymous biallelic sites that have shifted from OP in the ancestral codon to NOP.

<sup>c</sup> These include the derived synonymous codons that do not result in a shift among optimal and nonoptimal codons, that is, OP to OP and NOP to NOP.

<sup>d</sup> The number of derived codons at nonsynonymous polymorphic sites relative to ancestral codons.

### Frequency Distributions of Derived Optimal and Nonoptimal Codons

Based on the principles of the mutation–selection–drift model (Kimura 1983; McVean and Charlesworth 1999), advantageous derived optimal codons should segregate at higher frequencies within a population than deleterious derived nonoptimal codons (Akashi and Schaeffer 1997; Akashi 1999). In order to compare differences in average fitness effects of mutations, Sawyer et al. (1987) suggested that the frequency of two distinct categories of substitutions (herein, the categories are derived optimal versus derived nonoptimal codons) should be assessed by segregating each category of mutations into two frequency classes (FCs), namely “singletons” and a pool of all other FCs. Heterogeneity in the proportion of mutation categories found among these FCs constitutes evidence of differences in fitness effects (Sawyer et al. 1987; Akashi et al. 1999). Based on this principle, we determined the frequency spectra for derived optimal codons and derived nonoptimal codons in each N. crassa population examined herein (table 2). For this, we assessed the number of derived codons in each FC for the unfolded frequency spectrum (wherein the number of FCs \(N_{FC} = \) number of haplotypes \(-1\); thus, \(N_{FC \ Caribbean} = 18; N_{FC \ Louisiana} = 19\)). The results are provided in supplementary table S2, Supplementary Material online. We compared the proportion of derived optimal and nonoptimal codons in the highest FC for the Caribbean and Louisiana populations, that is, cases wherein all but one individual contained the derived codon (the “singletons”; FC 18 for the Caribbean population and FC 19 for Louisiana population). The results show that 314 of 5,399 derived optimal codon polymorphisms (5.8%) occur in the 18th FC whereas 475 of 11,341 derived
nonoptimal codons (4.18%) occur in this FC within the Caribbean population (supplementary table S2, Supplementary Material online), which is a statistically significant difference (chi-square test, degrees of freedom = 1, \( P < 0.0001 \)). Similar results were obtained for the Louisiana population, wherein 232 of the 5,740 derived optimal codons (4.04%) were found in the 19th FC as compared with only 356 of 12,094 (2.9%) derived nonoptimal codons (chi-square test, \( P < 0.0001 \)) (supplementary table S2, Supplementary Material online). Thus, the data provide evidence that derived optimal codons segregate at higher frequency than derived nonoptimal codons. A similar analysis was conducted for the lowest codon FC (FC = 1; wherein only one individual has the derived codon). The results show that the fraction of derived optimal codons represented in FC 1 is statistically significantly lower than the level of derived nonoptimal codons in both the Caribbean population (chi-square test, \( P < 0.0004 \)) and the Louisiana population (\( P < 0.0352 \)), that is consistent with a selective disadvantage of nonoptimal codons (Akashi and Schaeffer 1997). In sum, these data are consistent with a selective advantage of derived optimal codons and a selective disadvantage of nonoptimal codons in these \( N. \) crassa populations (Sawyer et al. 1987; Akashi 1999).

Another test for natural selection among derived optimal and derived nonoptimal codons is to assess whether the entire frequency distribution varies among these two types of codon substitutions using the Mann–Whitney U test (Akashi and Schaeffer 1997; Akashi 1999). Accordingly, one-sided Mann–Whitney U test was conducted for comparisons of the frequencies of optimal and nonoptimal codons as suggested by Akashi and Schaeffer (1997) and Akashi (1999) (note that two-sided tests yielded similar results). The analysis across all FCs show that even though derived optimal codons are less abundant (supplementary table S2, Supplementary Material online), they segregate at a higher frequency than derived nonoptimal codons in the Caribbean (Mann–Whitney U test, \( P < 10^{-15} \)) and in the Louisiana population (\( P < 10^{-15} \)), consistent with a selective advantage for optimal codons (fig. 1; supplementary table S2, Supplementary Material).

As a complementary analysis, we identified the derived optimal codons and derived nonoptimal codons that were specific to the Caribbean and the Louisiana population, which presumably represent the most recent mutations. For the Caribbean population, we identified a total of 1,232 (of 5,399, 22.8%) and 2,975 (of 11,341, 26.2%) population-specific derived optimal and derived nonoptimal codon mutations, whereas the respective values for the Louisiana population were 1,573 (of 5,740, 27.4%) and 3,728 (of 12,094, 30.8%). Analysis of this subset of mutations yielded similar trends to those obtained across all derived codons (supplementary fig. S1, Supplementary Material online), that is, derived optimal codon mutations segregate at a statistically significantly higher frequency than derived nonoptimal codon mutations in both the Caribbean population and the Louisiana populations (Mann–Whitney U test, \( P < 10^{-15} \), supplementary fig. S1, Supplementary Material online). Taken together, the frequency profiles of derived optimal and derived nonoptimal codons (fig. 1; supplementary table S2, supplementary fig. S1, Supplementary Material online) provide evidence of ongoing selection favoring optimal codon usage in both \( N. \) crassa populations (Akashi and Schaeffer 1997).

**Gene Length and Genetic Diversity**

The level of polymorphic sites relative to CDS length was assessed in order to more fully reveal the factors underlying optimal codon evolution in \( N. \) crassa. We assessed the relationship between the frequency of biallelic polymorphisms and gene length (based on the complete biallelic data set, table 1). For this, we determined the number of biallelic codon sites per 1,000 codons for every gene as follows: number of biallelic sites per 1,000 = number of biallelic codon sites in gene/number of codons in gene \( \times \) 1,000. The data show a highly statistically significant negative correlation between CDS length and the number of biallelic codon sites per 1,000 codons (fig. 2) for the data set.

---

**Fig. 1.** The frequency distribution of derived optimal (NOP to OP) and derived nonoptimal (OP to NOP) synonymous codons undergoing segregation at biallelic sites in \( Neurospora \) crassa. (A) Caribbean population. (B) Louisiana population.
from the Caribbean and for the Louisiana *N. crassa* populations (Spearman Rank Correlation \(R\) values ranging between \(0.32\) and \(0.51\), \(P < 10^{-15}\)). The correlation was found for both synonymous and nonsynonymous codon sites and indicates there is reduced genetic diversity in longer genes. This is consistent with genetic interference and/or genetic hitchhiking effects homogenizing biallelic codon sites in genes with extended CDS regions (Charlesworth et al. 1993; Comeron et al. 1999; Comeron and Guthrie 2005; Loewe and Charlesworth 2007; Bachtrog 2008; Ingvarsson 2010).

**Gene Length and Synonymous Codon Evolution**

Based on prior analyses at the macroevolutionary species level, gene (CDS) length has been inversely associated with bias in codon usage (e.g., Duret and Mouchiroud 1999; Cutter et al. 2008; Qiu et al. 2011a; Whittle et al. 2011b).

Thus, we assessed the association between gene length and codon usage at the microevolutionary level in these *N. crassa* populations. Specifically, we evaluated the relationship between gene length and the number of derived optimal and derived nonoptimal codons at biallelic codon sites (per 1,000 codons), across all genes with biallelic codon sites (table 1). Genes were placed into size bins of roughly equal frequency based on CDS length as follows: short (\(\leq 350\) codons), medium (\(> 350\) and \(\leq 600\) codons), and long (\(> 600\) codons). The numbers of genes categorized as short, medium, and long were 1,689, 1,779, and 1,776 in the Caribbean data set and 1,652, 1,760, and 1,770 in the Louisiana data set, respectively. The detection rate of ancestral codons was nearly identical among gene lengths. The results show that the level of derived optimal and derived nonoptimal codons at biallelic sites (per 1,000 codons of gene length) was higher for longer than shorter genes (supplementary fig. S2, Supplementary Material online); this is in agreement with the reduced frequency of biallelic codon sites in longer genes (fig. 2). Accordingly, it may be inferred that the reduced frequency of biallelic codon sites in long genes translates into reduced levels of derived optimal and derived nonoptimal codons (per 1,000 codons) in these *N. crassa* populations.

We attribute the differences in the frequency of biallelic codon sites and derived optimal and derived nonoptimal codons (per 1,000 codons) (fig. 2; supplementary fig. S2, Supplementary Material online) to selective variation among gene length categories. This is because prior data have shown that selection, and not mutation, drives optimal codon usage across genes of various lengths in certain species of *Neurospora* (Whittle et al. 2011b). We further evaluated this relationship herein for *N. crassa* as follows.

**Fig. 2.** The number of biallelic codon sites per 1,000 codons for each of the genes examined in the present study. (A) Synonymous codon sites in the Caribbean population; (B) nonsynonymous codon sites in the Caribbean population; (C) synonymous codon sites in the Louisiana population; and (D) nonsynonymous codon sites in the Louisiana population. Spearman rank correlation coefficient \(R\) values and \(P\) values are shown. A small number of extreme values of gene length (less than eight genes) and genes with values of zero (genes lacking a synonymous biallelic site or lacking a nonsynonymous biallelic site) have been excluded from the figures.
Given that the majority of primary optimal codons (the codon most favored in highly expressed genes per amino acid) have G or C at the third nucleotide position (GC3) (supplementary table S1, Supplementary Material online), we compared the GC3 content to the GCI (which are selectively neutral or nearly selectively neutral) for all three gene-length categories (sample size of introns in each gene length category: short N = 1,452, medium N = 1,453, long N = 1,480). We note that all genes having introns (among genes having SNPs and full CDS in N. crassa) were examined, regardless of whether they contained SNP's in the Caribbean and Louisiana populations or both populations.

The data show that, consistent with prior findings for N. tetrasperma and N. discreta, GC3 is statistically significantly higher than GCI for short, medium, and long genes (supplementary fig. S3, Supplementary Material online, Wilcoxon paired ranked sum test P < 10^-15 for each gene length). This confirms that selection and not mutation primarily drives optimal codon usage across the different gene length categories in N. crassa. In addition, the GC3 content for CDS regions (an indicator of optimal codon usage) is highest in short genes, intermediate in medium-length genes, and lowest in long genes, which is consistent with reduced optimal codon usage with greater gene lengths (Mann–Whitney U test, P < 10^-15 for all comparisons, supplementary fig. S3, Supplementary Material online). This finding further supports the notion that the lower level of biallelic codon sites (fig. 2) and derived optimal codons (per 1,000 codons) in longer genes (supplementary fig. S2, Supplementary Material online) in N. crassa populations examined herein are best explained by genetic interference; such a phenomenon is consistent with susceptibility to genetic hitchhiking in longer genes (Charlesworth et al. 1993; Comeron et al. 1999; Comeron and Guthrie 2005; Loewe and Charlesworth 2007; Bachtrog 2008; Ingvarsson 2010).

**Optimal Codons versus Nonsynonymous Substitutions**

Although gene length effects are consistent with genetic-linkage effects at the population level, another means that assess a long-term history of genetic-linkage effects on optimal codon usage in N. crassa is to compare protein evolution and the Fop. Prior data in Drosophila showing an inverse correlation between protein evolution and Fop have revealed that adaptive protein changes interfere with selection at linked sites, including optimal codons (Betancourt and Presgraves 2002; Shapiro et al. 2007; Bachtrog 2008). Thus, we assessed the relationship between these parameters for N. crassa (using N. tetrasperma as the reference taxon and across genes with the full CDS identified in alignments for all three Neurospora taxa; see Materials and Methods, Ngenes = 1,643). For this, N. tetrasperma was used as the outgroup taxon due to its close relatedness to N. crassa (Dettman et al. 2003); however, N. discreta yielded highly similar results. Fop values were determined using CodonW (Peden, http://codonw.sourceforge.net/). The data show that there is a highly statistically significantly negative correlation between the number of nonsynonymous codon substitutions (per 1,000 codons) and the Fop in N. crassa (Spearman rank correlation [R] = −0.30, P < 10^-15, supplementary fig. S4, Supplementary Material online); this suggests that adaptive protein evolution interferes with selection on codon usage at linked sites in this taxon. In addition, we evaluated the correlation for genes from various gene length categories. We found a statistically significant correlation between nonsynonymous codon substitutions and Fop for short, medium, and long genes (supplementary fig. S4, Supplementary Material online). Notably, the long genes had the largest (negative) correlation coefficient (R = −0.38, P < 10^-15) among all three gene-length categories, which is nearly twice the value for short genes (R = −0.20, P < 10^-15). This trend suggests that genetic hitchhiking in rapidly evolving genes has its greatest effect in longer genes. In this regard, these data (supplementary fig. S4, Supplementary Material online), combined with our results showing reduced polymorphism and derived optimal codons (fig. 2; supplementary fig. S2, Supplementary Material online) and lowered levels of optimal codons in longer genes (supplementary fig. S3, Supplementary Material online), suggest that genetic hitchhiking interferes with optimal codon usage in N. crassa. Accordingly, the totality of the data indicates that gene length is a significant factor driving optimal codon frequencies at the population level (observed in fig. 2; supplementary fig. S2, Supplementary Material online) and in their ultimate fixation (supplementary figs. S3 and S4, Supplementary Material online).

**Frequency Distributions of Derived Optimal Codons Relative to Gene Length**

Given that longer genes show evidence of genetic interference, we assessed whether the selective advantage of optimal codons relative to nonoptimal codons in each N. crassa population examined herein is dependent on gene length. Specifically, we determined the frequency distributions of derived optimal codons and derived nonoptimal codons in each population for the three gene length categories. Similar to the findings across all genes in fig. 1, we found that the derived optimal codons segregate at a higher frequency than derived nonoptimal codons for short, medium, and for long genes (fig. 3; Mann–Whitney U test, P < 0.05 for each of the three gene length categories) in both the Caribbean and Louisiana populations; thereby revealing that this trait is not gene-length dependent. Thus, even though the number of biallelic codon sites and the level of derived optimal and derived nonoptimal codons (per 1,000 codons) are each reduced in long genes (fig. 2; supplementary fig. S2, Supplementary Material online), this reduction does not alter the trend of elevated segregation frequencies for derived optimal codons as compared with derived nonoptimal codons in the population (fig. 3). The lack of gene length effects on the selective advantage of optimal codons is further supported by the fact that the frequency distributions of derived optimal codons as well as the frequency distributions of derived nonoptimal codons remained similar across all gene length categories.
codons were similar among short, medium, and long genes (P values were between 0.17 and 0.97 for all pairwise comparisons among the various gene lengths using two-sided Mann–Whitney U test, fig. 3). These data show that the selective advantage of derived optimal codons and selective disadvantage of derived nonoptimal codons is maintained across all gene length lengths, including long genes. Taken together, it may be inferred that derived optimal codons retain selective advantage relative to derived nonoptimal codons in longer genes in the *N. crassa* populations (fig. 3), but the standing polymorphism level for derived optimal codons and nonoptimal codons is lowered in longer genes (fig. 2; supplementary fig. S2, Supplementary Material online) due to genetic linkage effects (supplementary figs. S3 and S4, Supplementary Material online).

**Frequency Distribution of Nonsynonymous Codon Mutations**

We assessed the frequency distributions of derived nonsynonymous mutations within each *N. crassa* population. For all comparisons of frequency distributions involving nonsynonymous mutations, two-sided tests were conducted as no a priori assumption was made on the direction of the difference. The frequency distributions of derived nonsynonymous codons across *N. crassa* biallelic sites reveal that the vast majority of these mutations segregate at low frequencies in the populations (fig. 4), consistent with substantial deleterious effects of these mutations. The derived optimal codons as well as the derived nonoptimal codons (which segregate at a lower frequency than optimal codons, fig. 1) each segregate at higher frequencies than derived nonsynonymous codons in both populations (Mann–Whitney U test have P < 10^{-15} for each comparison per population). Thus, it may be concluded that the frequency of segregating mutations in each *N. crassa* population decreases in the following manner: synonymous derived optimal codons > derived nonsynonymous codons. In totality, these data are consistent with greater selective pressure against nonsynonymous codon mutations than synonymous mutations in each *N. crassa* population.

**Discussion**

Selective differences among two distinct classes of mutations in a population may be exposed by comparing their frequency distributions at polymorphic sites (Sawyer et al. 1987; Tajima 1989; Akashi and Schaeffer 1997). In particular, comparison of frequency spectra of derived optimal and nonoptimal codons at the population level controls for the evolutionary history of the of sample of DNA sequences, including variation in population sizes or mutation rate over time (Sawyer et al. 1987; Akashi and Schaeffer 1997; Akashi 1999). The power of such analyses to detect distinct fitness classes of mutations is increased by using ancestral codons to determine the direction of

---

**Fig. 3.** (A–B) The frequency of derived optimal and nonoptimal codons in the *Neurospora crassa* populations for short, medium, and long genes. Different letters within a single gene length category indicate a statistically significant difference using the Mann–Whitney U test. The proportion of the sample (y axis) shown was determined across all codon sites and represents the mean number of individuals containing a derived optimal (or nonoptimal) codon mutation/total number of individuals per sample (N = 19 for the Caribbean population and N = 20 for the Louisiana population).

**Fig. 4.** The frequency distribution of nonsynonymous derived mutations in the Caribbean and Louisiana populations.
mutation (e.g., ancestral or derived optimal codons) in a precise manner (Akashi and Schaeffer 1997). Herein, the criterion for defining ancestral codons and the direction of optimal codon mutations in N. crassa was stringent and was based on comparison to its two close relatives N. tetrasperma and N. discreta. Our findings of skewed frequency of derived optimal codons toward the highest FCs and of derived nonoptimal codons toward the lowest FCs (fig. 1) provides marked evidence of selection favoring optimal codons and disfavoring nonoptimal codons within the N. crassa Caribbean and Louisiana populations.

Although studies of derived optimal codons at polymorphic sites have been rare, data from Drosophila and Silene have supported similar patterns of selective pressure on codon usage in populations. Comparative analyses of derived optimal and nonoptimal codons from D. simulans, D. pseudoobscura, D. mauritiana, and D. americana, and D. melanogaster and S. latifolia (across 18 or fewer loci) are each consistent with the phenomenon of selection favoring optimal (or preferred) codon usage within their populations (Akashi and Schaeffer 1997; Maside et al. 2004; Llopart et al. 2008; Qiu et al. 2011a). Analysis of codon usage patterns in D. melanogaster across 468 loci and in Arabidopsis and Capsella spanning 780 and 354 exon loci respectively, based on an alternate technique than used herein, are also consistent with selection favoring optimal codons at the population level (Zeng and Charlesworth 2009; Qiu et al. 2011b). In the present analysis, spanning more than 5,100 genes, we provide evidence that differential selection on optimal and nonoptimal codons occurs at a genome-wide scale in N. crassa populations.

Factors Underlying Selection on Codon Usage

Selective pressure on codon usage is believed to be weak, with the product of NeS ≈ 1 (Akashi and Schaeffer 1997; McVean and Vieira 2001). Thus, selection is believed to be most efficient, relative to genetic drift, in organisms with large population size (Ikemura 1985; Sharp and Li 1986). The variation in frequency distributions for derived optimal and nonoptimal codons reported herein suggests that the size of the Caribbean and Louisiana populations is sufficient to support effective selection. A large effective population size in N. crassa is facilitated by the high level of outbreeding in this heterothallic fungus (Powell et al. 2003; Ellison et al. 2011), which enhances population size compared with inbreeding (Charlesworth and Wright 2001; Whittle et al. 2011). In addition, a large population size for this fungal taxon is consistent with the reported wide-scale distributions for the Louisiana and Caribbean populations. For example, isolates from the Caribbean population span Florida, Haiti and the Yucatan. In this regard, population level selection on codon usage may be particularly evident in the present investigation owing to extensive population sizes in N. crassa.

Another key factor that likely contributes toward the skewed frequency of derived optimal codons in the higher FCs in N. crassa is haploid selection. N. crassa spends the vast majority of its lifespan as a haploid; the diploid stage is restricted to the short zygote stage (Shear and Dodge 1927; Shiu and Glass 2000), and accordingly nonoptimal derived codons are not sheltered by a homologous nonmutant allele as occurs in diploids/polyplids. Haploid selection is believed to limit fixation of nonpreferred codon mutations and nonsynonymous substitutions in regions of suppressed recombination in the mat chromosome regions in N. tetrasperma (Whittle and Johannesson 2011; Whittle et al. 2011a). In addition, it has been inferred that haploid selection contributes toward the restriction of genomic degeneration (e.g., gene losses) in other species of fungi such as Cryptococcus neoformans due to the absence of mutational sheltering (Fraser et al. 2004). These observations suggest that the finding of ongoing selection for optimal codon usage in N. crassa is facilitated by the marked opportunity for haploid selection within this taxon.

Gene Length and Derived Optimal Codons

Theory suggests that genetic hitchhiking (Hill–Robertson effects) interferes with the effectiveness of selection in a manner that is dependent on gene length, even for genomic regions with recombination (Comeron and Guthrie 2005; Andolfatto 2007; Loewe and Charlesworth 2007). Prior findings showing that codon usage bias tends to be reduced for longer genes at the species level in various organisms, including species of Neurospora, may be due to such genetic linkage effects (Duret and Mouchiroud 1999; Comeron and Guthrie 2005; Loewe and Charlesworth 2007; Whittle et al. 2011b). In turn, such processes also likely influence optimal codon mutations observed within the N. crassa populations examined herein (Charlesworth et al. 1993; Comeron and Guthrie 2005). In particular, the findings that longer genes have markedly lower levels of biallelic sites for N. crassa (per 1,000 codons, fig. 2) is consistent with genetic hitchhiking converting many biallelic sites into homoallelic sites (Kaplan and Hudson 1989; Charlesworth et al. 1993; Wiehe and Stephan 1993). This also explains the lower frequency of derived optimal and derived nonoptimal codons found in longer genes (per 1,000 codons, supplementary fig. S2, Supplementary Material online). In addition, genetic hitchhiking in N. crassa is also supported by the reduced levels of fixed optimal codons (GC3) in longer genes (supplementary fig. S3, Supplementary Material online). Similar findings have been reported in species of Drosophila and Populus wherein genetic hitchhiking has been associated with reduced standing polymorphism levels and/or lower optimal codon frequency within genes (Betancourt and Presgraves 2002; Andolfatto 2007; Bachtrog 2008; Ingvason 2010). The trend of reduced genetic variation at the population level may explain why fixation of optimal codons tends to be lower in longer genes at the macroevolutionary scale in species of Neurospora (supplementary fig. S3, Supplementary Material online; Whittle et al. 2011b), as well as other organisms (Duret and Mouchiroud 1999; Cutter et al. 2008; Qiu et al. 2011a).

In addition to our findings that longer genes have reduced genetic diversity, the role of Hill–Robertson effects in shaping optimal codon usage is also consistent with the fact that the Fop was inversely correlated with the level of protein
evolution in *N. crassa* (supplementary fig. S4A, Supplementary Material online). Such a relationship suggests that protein changes are influencing codon usage at linked sites (Betancourt and Presgraves 2002; Bachtrog 2008; Ingvarsson 2010). Prior studies in Drosophila have shown a similar inverse correlation between protein evolution and codon usage, which has been attributed to recurrent adaptive protein evolution events dragging linked codon mutations to fixation via selective sweeps (Betancourt and Presgraves 2002; Andolfatto 2007; Bachtrog 2008). Our present data reveal that the strength of this correlation is gene-length dependent and is markedly enhanced in longer genes in *N. crassa* (supplementary fig. S4B,C,D, Supplementary Material online). Taken together, the present data suggest that long genes, and particularly rapidly evolving long genes, are highly vulnerable to reduced fixation of optimal codons due to interference by genetic hitchhiking in *N. crassa*.

Despite the fact that long genes have fewer biallelic sites and derived optimal codons per 1,000 codons (fig. 2; supplementary fig. S2, Supplementary Material online), our findings show that derived optimal codons are skewed toward the upper FCs as compared with derived nonoptimal codons across all three categories of gene length, including long genes (fig. 3). Accordingly, it may be inferred that the selective advantage of optimal codons is robust to gene length in these *N. crassa* populations. It may be speculated that the remaining derived optimal and derived nonoptimal codons at biallelic codons sites, that is, those not eliminated by genetic hitchhiking effects, are subjected to similar selective pressures for optimal codon usage as shorter genes. In this regard, the fixation of optimal codons in long genes is likely dependent on the balance between genetic hitchhiking and active selection on optimal codons. In the future, simulation studies (e.g., Akashi 1999) that model the dynamics between genetic linkage effects and selection favoring optimal codon usage in populations will be useful to more fully understand how these parameters interact to ultimately determine the fixation rate of optimal codons in long genes.

**Nonsynonymous Codon Mutations**

The findings that derived nonsynonymous codon mutations segregate at very low frequency in *N. crassa* populations are consistent with a deleterious effect of a substantial fraction of amino acid changes (Akashi 1996). Findings based on polymorphism data in populations of *D. simulans* and *D. melanogaster* and from *Escherichia coli* also support elevated segregation frequencies for synonymous mutations relative to nonsynonymous mutations (Sawyer et al. 1987; Akashi and Schaeffer 1997). Our findings showing that nonsynonymous mutations segregate at lower frequency than synonymous derived nonoptimal codons suggests that the average deleterious effects of amino acid changes exceed that of mutations to nonoptimal codons in *N. crassa* (Akashi 1996). This is in agreement with the fact that many amino acid changes have detrimental fitness effects due to alterations in critical cellular pathways and physiological processes (via alternations in protein function, e.g., Ng and Henikoff 2006; Larracuent et al. 2008) whereas deleterious synonymous nonoptimal codon mutations primarily influence translational efficiency (that presumably on average has less severe consequences, Akashi 1999, 2001). The fact that the frequency of segregating mutations decreases in the following order: synonymous derived optimal codons, synonymous derived nonoptimal codons, and derived nonsynonymous codons is consistent with a progressive increase in the deleterious effects of these categories of mutations in *N. crassa*.

**Notable Issues**

It is worthwhile to note that our data set of biallelic codon sites is based on those *N. crassa* codon sites wherein an ancestral codon was identified (table 2), which could conceivably differ slightly from a random and/or unfiltered data set of all possible codon sites. For instance, given that the ancestral codons are those that are identical between *N. tetrasperma* and *N. discreta*, the current data set is based on relatively conservative codon sites, which could have a different potential for codon mutations and/or selective pressure than across all polymorphic codon sites. In addition, it is notable that each of the populations may not be at equilibrium (fig. 1, Zeng and Charlesworth 2009). This is not apt to be consequential to our findings as our methodology accounts for population demographics (Hudson 1993; Akashi and Schaeffer 1997; Akashi 1999). Furthermore, both populations show identical trends consistent with a selective advantage of derived optimal over derived nonoptimal codons. Specifically, the frequency distributions are shaped in a manner precisely matching a selective advantage of derived optimal codons (fig. 1), and the pattern is identical for short, medium, and long genes (fig. 3). Taken together, these factors argue against a significant role of population demographics in the present study.

**Conclusion**

The present analyses based on large-scale SNP data sets in two distinct *N. crassa* populations reveals that selection favors derived optimal codons at segregating codon sites. These findings support ongoing selection for optimal codon usage in *N. crassa*. In addition, the data reveal that while gene length plays a significant role in genetic interference, and thus the level of genetic variation in the population, selection favoring optimal codon usage is retained across all gene lengths. Further, the data show that nonsynonymous changes have on average greater deleterious effects than derived optimal and derived nonoptimal codon mutations, consistent with detrimental effects on cell pathways and physiology. Additional studies will be needed to reveal whether these phenomena are evident in other species of Neurospora, particularly those with inbreeding mating systems and various population sizes in order to further reveal the types of factors driving evolution of codon usage at the population level for this taxonomic group.
Supplementary Material

Two supplemental tables and four figures are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors gratefully acknowledge research funding from the Erik Philip-Sörensens Stiftelse (C.A.W.), the Lars Hiertas Foundation (C.A.W.), and the Swedish Research Council (H.J.). We also thank S. Otto, P. Corcoran, and two anonymous reviewers for valuable comments on the manuscript.

References


