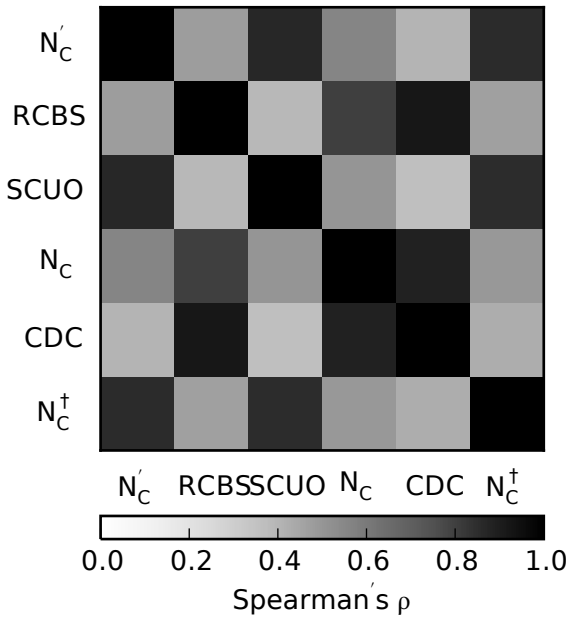
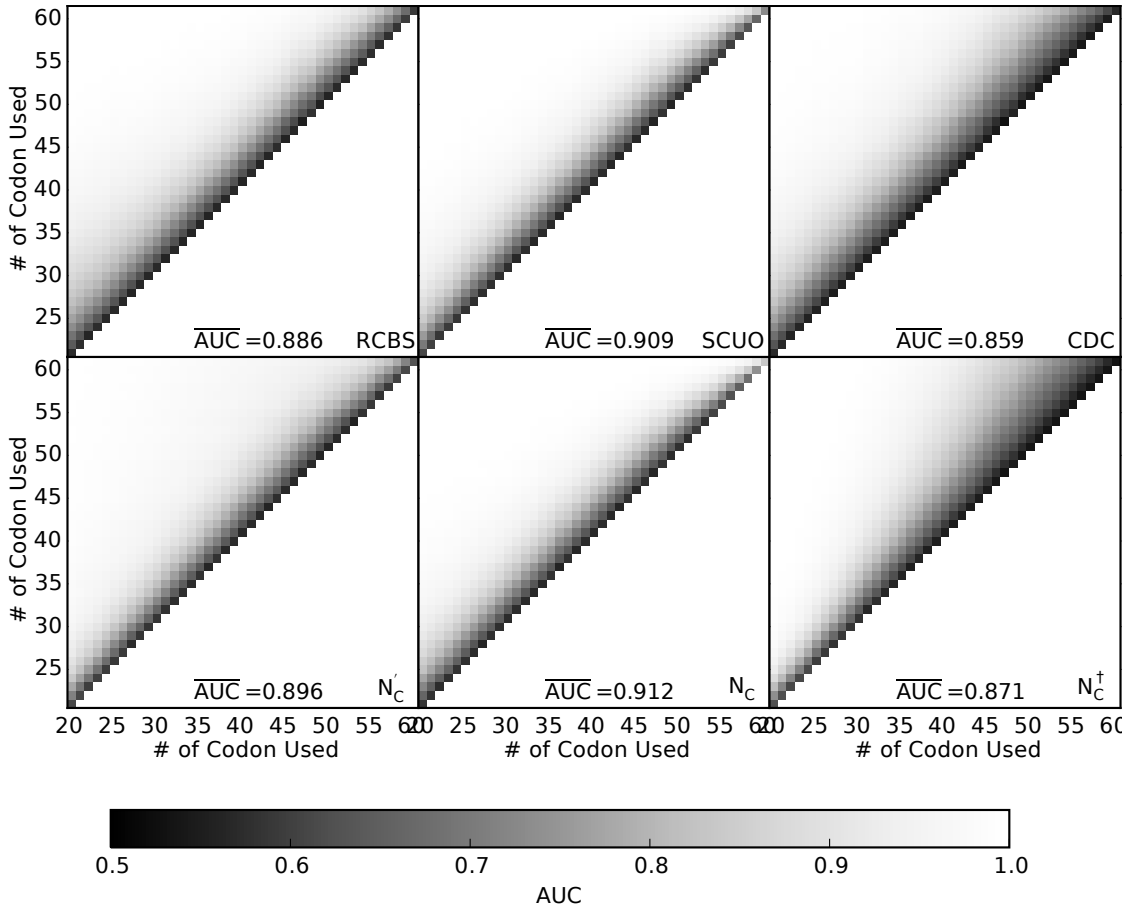
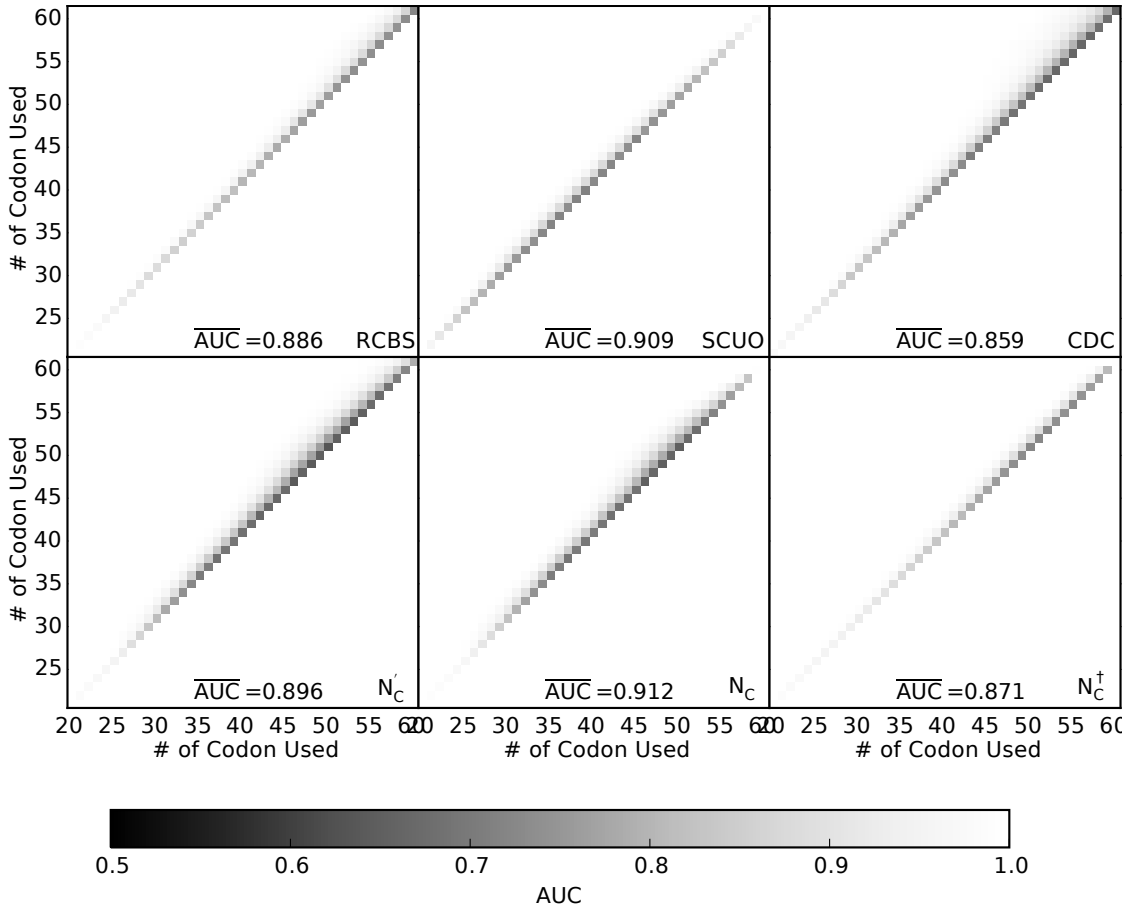
Supplementary Information



Supplementary Figure 1. Spearman's ρ between the calculated CUB of 3740 **genes in the *E. coli* genome using various metrics.** All correlations are statistically significant with $p < 10^{-10}$



Supplementary Figure 2. Performance of individual pair-wise differentiation task for six different CUB metrics for one bootstrapped sample using *E. coli*-like genome. Each block in the heatmap represents the AUC when using that metric to differentiate sequences created using different number of codons.



Supplementary Figure 3. Performance of individual pair-wise differentiation task for six different CUB metrics for one bootstrapped sample under the simplest case. Each block in the heatmap represents the AUC when using that metric to differentiate sequences created using different number of codons.

Supplementary Table 1. P-value of the Wilcoxon signed-rank test for gene expression correlations between different metrics. Bold face values indicate when the test statistic is positive (i.e. when the metric listed at the left is higher than the metric listed at the top).

Metric	N'_C	N_C	$N^{\dagger}_C$	RCBS	SCUO	CDC
N'_C	n/a	**<0.001**	**<0.001**	**0.124**	**0.086**	**0.018**
N_C		n/a	**0.328**	0.485	0.209	0.929
$N^{\dagger}_C$			n/a	0.328	0.078	0.751
RCBS				n/a	0.909	**0.022**
SCUO					n/a	**0.03**
CDC						n/a

SI text: Implementation of previous CUB metrics

Supplementary Table 2. Definitions of variables

Variable	Definition
A_i	i-th amino acid in a given translation table
c_{ij}	j-th synonymous codon of the A_i
N_{ijl}	the l-th nucleotide of c_ij
p_{ij}	probability of picking c_{ij}
$p(c_{ij}\|A_i)$	probability of picking c_{ij} given A_i
$p(A_i)$	probability of observing the amino acid that codes for c_ij
$p(N_{ijl})$	the observed probability of N_{ijl}
$f_x(N_{ijl})$	the observed probability of N_{ijl} in the x-th nucleotide position of all codons in the sequence
k_i	number codons that code for A_i
K	total number of available sense codons, $K = \sum_{i=1}^{20} k_i$
k_{A_i}	number of degenerate codons that code for A_i
n_{A_i}	number of times A_i is observed in the sequence
$n_{c_{ij}}$	number of time c_{ij} observed in the sequence
L	number of codons in the sequence

We benchmarked six different CUB metrics (N_C, N'_C, $N_C^{\dagger}$, SCUO, RCBS, and CDC). All calculations were performed according to the original papers.

Calculating N_C. The CUB metric N_C is calculated from F of an amino acid (F_{A_i}), also known as homozygosity which is defined as:

$$F_{A_i} = \frac{n_{A_i} \sum_{j=1}^{k_i} p(c_{ij}|A_i)^2 - 1}{n_{A_i} - 1} \tag{1}$$

Once F_{A_i} of each amino acid determined, the CUB of the gene is determined by:

$$N_C = 2 + \frac{9}{\bar{F}_2} + \frac{1}{\bar{F}_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6} \tag{2}$$

where $\bar{F}_d$ is the average of all F_{A_i} in the d-fold degeneracy class. In the case where the observed frequency of a particular amino acid AA is 0, F_{A_i} will be excluded from the calculation of $\bar{F}_d$. If $\bar{F}_3$ can not be calculated, $\bar{F}_3$ will be the average of $\bar{F}_2$ and $\bar{F}_4$. If for a j-fold degeneracy class $\bar{F}_d$ can not be calculated, then $\bar{F}_d = \frac{1}{d}$. In the method, all sequences with a calculated N_C greater than 61 will be readjusted to 61.

Calculating N'_C. The CUB metric N'_C is calculated from F of an amino acid, which is defined as:

$$F_{A_i} = \frac{\chi^2_{A_i} + n_{A_i} - k_{A_i}}{k_{A_i}(n_{A_i} - 1)} \tag{3}$$

where $\chi^2_{A_i}$ is the χ^2 statistic for amino acid AA which is given by the equation:

$$\chi^2_{A_i} = \sum_{j=1}^{k_i} \frac{n_{A_i}\big(p(c_{ij}|A_i) - \widehat{p}(c_{ij}|A_i)\big)^2}{\widehat{p}(c_{ij}|A_i)} \tag{4}$$

where $\widehat{p}(c_{ij}|A_i)$ is the expected usage of the codon c_{ij} given the observed nucleotide probabilities in the sequence, described by the equation:

$$\widehat{p}(c_{ij}|A_i) = \frac{1}{Z}\prod_{l=1}^{3} p(N_{ijl}) \tag{5}$$

where Z is a normalization factor so that $\sum_{j=1}^{k_i} \widehat{p}(c_{ij}|A_i) = 1$ for all amino acids.

From here the final CUB value is calculated exactly the same as in equation 2. In the case where the observed frequency of a particular amino acid A_i is 0, F_{A_i} will be excluded from the calculation of $\bar{F}_d$. If $\bar{F}_3$ can not be calculated, $\bar{F}_3$ will be the average of $\bar{F}_2$ and $\bar{F}_4$. If for a d-fold degeneracy class $\bar{F}_d$ can not be calculated, then $\bar{F}_d = \frac{1}{d}$. Amino acids where less than 5 occurrences is observed are excluded. In the method, all sequences with a calculated N'_C greater than 61 will be readjusted to 61.

Calculating $N_C^\dagger$. In this method, 6-fold degenerate amino acids are broken up into a two-fold and a four-fold degenerate amino acid. That is to say for example, in the case of Leucine (TTA, TTG, CTT, CTC, CTA, CTG), TTA and TTG will be group together as a two-fold degenerate amino acid, and CTT, CTC, CTA, and CTG will be grouped together as a four-fold degenerate amino acid. The CUB of a sequence is:

$$N_C^\dagger = D_1 + \sum_{d=2}^{4} \frac{D_d \sum_{x=1}^{D_d} n_{A_{dx}}}{\sum_{x=1}^{D_d} n_{A_{dx}} F_{A_{dx}}} \tag{6}$$

where D_d is the number of d-fold degenerate amino acids, A_{dx} is the x-th amino acid in a set of d-fold degenerate amino acids, and F_{A_i} is the homozygosity of amino acid A_i defined by:

$$F_{A_i} = \sum_{i=1}^{20} \left(\frac{n_{c_{ij}} + 1}{n_{A_i} + k_i}\right)^2 \tag{7}$$

Calculating SCUO. The synonymous codon usage order (SCUO) is an informatics based method in which the CUB of a sequence is calculated from the entropy of the sequence. In the method, the entropy of an amino acid A_i is defined by:

$$H_{A_i} = -\sum_{j=1}^{k_i} p(c_{ij}|A_i) \log_2 p(c_{ij}|A_i) \tag{8}$$

An observable difference in the entropy of the sequence and maximal entropy can be expressed as:

$$O_{A_i} = \frac{H_{A_i}^{max} - H_{A_i}}{H_{A_i}^{max}} \tag{9}$$

where $H_{A_i}^{max} = -\log k_{A_i}$. To obtain the total CUB of the entire sequence, the individual observable difference of each amino acid is aggregated by:

$$SCUO = \sum_{i=1}^{20} p(A_i) O_{A_i} \tag{10}$$

Calculating RCBS. The relative codon bias score (RCBS) of a sequence is defined as:

$$RCBS = \prod_{i,j} (1 + d_{c_{ij}})^{\frac{n_{c_{ij}}}{L}} - 1 \tag{11}$$

where d_C is the difference between the expected and actual codon probability observed for the sequence. This is defined by:

$$d_C = \frac{p(c_{ij}|A_i) - \widehat{p}(c_{ij}|A_i)}{\widehat{p}(c_{ij}|A_i)} \tag{12}$$

where $\widehat{p}(c_{ij}|A_i)$ is defined by:

$$\widehat{p}(c_{ij}|A_i) = \frac{1}{Z} \prod_{l=1}^{3} f_l(N_{ijl}) \tag{13}$$

where Z is a normalization factor so that $\sum_{j=1}^{k_i} \widehat{p}(c_{ij}|A_i) = 1$.

Calculating CDC. The CUB metric CDC compares the bias of a gene based on nucleotide content of the sequence. CDC is defined by the following equation:

$$CDC = \frac{\sum_{i=1}^{20} \sum_{j=1}^{k_i} p_{ij} \widehat{p}_{ij}}{\sqrt{\left(\sum_{i=1}^{20} \sum_{j=1}^{k_i} p_{ij}^2\right)\left(\sum_{i=1}^{20} \sum_{j=1}^{k_i} \widehat{p}_{ij}^2\right)}} \tag{14}$$

where $\widehat{p}_{ij}$ is the expected probability of codon c_{ij}. The expected probability of codon c_{ij} is defined by the expected nucleotide probability at each nucleotide position in the codon, which is then defined based on the GC and Guanine content at each nucleotide position in the codon. The expected probability of each nucleotide at the lth nucleotide position in the codon is defined in the following four equations:

$$a_l = (1 - S_l)R_l \tag{15}$$

$$t_l = (1 - S_l)(1 - R_l) \tag{16}$$

$$g_l = S_l R_l \tag{17}$$

$$c_l = S_l(1 - R_l) \tag{18}$$

where S_l and R_l are the GC and purine content at lth nucleotide position of the codon respectively. $\widehat{p}_{ij}$ is then define by:

$$\widehat{P}_{ij} = \frac{x_1 y_2 z_3}{Z} \tag{19}$$

where $x, y, z \in \{A, T, G, C\}$ and Z is a normalization factor so that $\sum_{i=1}^{20} \sum_{j=1}^{k_i} \widehat{p}_{ij} = 1$.