Changes in codon usage upon stress are not explained by changes in either amino acid usage or nucleotide usage

Having established that the codon usage changes dynamically during stress we wished to examine whether the change in the representation of a given codon can be explained by a corresponding change in the representation in the transcriptome of the nucleotides that constitute that codon, or alternatively by a change in the representation of its respective amino acid in the translated transcriptome. To examine these two alternative hypotheses we computed the nucleotide and amino acid expression matrices under the same stress conditions. The "Nucleotide expression matrix" is a 4xN matrix whose i,j-th element indicates the extent of appearance of nucleotide i in the transcriptome at condition or time point j. The "Amino acid expression" matrix is a 20xN matrix whose i,j-th element depicts representation of amino acid i at the translated transcriptome at condition or time point j. With the nucleotide expression matrix we ask whether changes at the codon expression matrix can be reduced to, and explained by, changes at the representation of the various nucleotides. Such changes may be related to putative changes in the nucleotide composition of the transcriptome.(1). Likewise, the amino acid expression matrix allows us to ask whether changes at the codon expression matrix simply reflect changes in the relative appearance of the different amino acids at the translated transcriptome, changes that my occur in specific amino-acid cases (2).

We detected only moderate fluctuations in the usage of amino acids upon stresses compared to the changes in the usage of individual codons (Figure S2). Is it possible that the changes in codon usage are simply derived from these changes in amino acid usage? For this purpose we calculated the partial correlations between fold-changes in the representation of individual codons upon stress and the translational efficiency (by the tAI measure) of these codons, while controlling for fold-changes in the usage of the respective amino acids. This analysis shows at most a negligible effect of variations in consumption of different amino acids on the preference of low-efficiency codons upon stress (all partial correlations are very close to the original correlation values). Using the "Nucleotide-Expression" matrix, we detected slight fluctuations in the GC content of the transcriptome upon different types of stress – fold-changes values vary between 0.99-1.01 and 0.98-1.03 for codon position-independent and codon position-dependent usage of nucleotides, respectively (Figure S2).

Exploring the balance between drift and selection by a computational simulation

We developed a computer simulation of a simplified evolutionary process of unicellular population of a fixed size of 1,000,000 haploid cells for 10,000 generations. The genome of each cell consists of six genes – a house-keeping gene that is expressed in every environment and growth condition, a 'good-life' gene, corresponds to favorable growth conditions, three 'stress-specific' genes, which are uniquely associated with three different stress types, and a 'stress-generic' gene, which is essential for any stress type.

At the beginning of the simulation, the six genes are equally scored with initial arbitrary value of expression level that denotes optimal expression. The population then evolves while subjected to a fixed mutation rate, that is, the frequency of 0.001 substitutions per genome, in line with realistic values (3,4). Sequences are not represented explicitly in the simulation; instead genes are characterized by an expression level that implicitly corresponds to a genotype. Thus, "mutated" expression levels at a given time step are computed by the previous step's expression levels multiplied by a random number drawn from an exponential probability distribution of changes in expression (as estimated before (REF 5)).

We set the rate parameter λ to be 1.5, hence approximately eighty percent of the mutations are assumed to be deleterious. Running the simulation with less deleterious mutations ($\lambda = 1$), reproduces the results.

We ran the simulation in two modes. In the first, mutations affected the expression of genes, but there was no bound on the total expression level for all the genes in the genome. In the second mode of the simulation mutations affected expression as in the first mode, yet in addition the tRNAs supply is limited, so that not all genes can be optimized simultaneously. Practically, we forced a constant maximal total expression level from all genes. In this mode of limited supply of tRNAs, the expression of the *i*-th gene in each generation, $lsExpression_{gi}$, ("ls" stands for "limited supply") is defined by

$$lsExpression_{gi} = \begin{cases} Expression_{gi} * \left(\frac{MaxExpression}{\sum_{i=1}^{n} Expression_{gi}} \right) & if \left(\sum_{i=1}^{n} Expression_{gi} > MaxExpression \right) \\ Expression_{gi} & else \end{cases}$$

where n is the number of genes in the cell, $Expression_{gi}$ is the expression of the i-th gene in the mutated population, and $MaxExpression$ is a constant maximal total expression level from the whole "genome".

The evolving population of cells is exposed to occasional stress periods that come in three types, stress1, stress2 and stress3. Specifically, we applied three different regimes, in which the total duration of stressful conditions constitutes 20, 50 or 80 percent of the total evolutionary time.

Individual cells are selectively transferred for the next generation, as a function of their fitness. The fitness of a given cell is determined by a weight given to it according to the expression of its genes which are associated with the current environmental condition during which a distinct cell division event occurs. Specifically, the fitness in favorable growth conditions is a function of the expression values of the 'house-keeping' and 'good-life' genes, whereas the fitness during stress is determined as the averaged expression value of the 'house-keeping' gene, the relevant 'stress-specific' gene and the general stress gene. Practically, we measured the change in the fitness of individual cells as the absolute value of the difference of expression values of the condition-related genes from the optimal one. Having the fitness values for all the cells in the population, the simulation program selects cells for the next generation. Formally, the numeric change in the size of homogeneous population can be described as

$$\frac{dx}{dt} = \lambda x\left(1 - \frac{x}{K}\right)$$

where x denotes the population size, λ corresponds to the fitness, and K indicates the carrying capacity according to the logistic model. For a heterogeneous population consisting of two genotypes, the respective equations are

$$\frac{dx_1}{dt} = \lambda_1 x_1\left(1 - \frac{x_1 + \alpha_{21}x_2}{K}\right) \text{ and } \frac{dx_2}{dt} = \lambda_2 x_2\left(1 - \frac{x_2 + \alpha_{12}x_1}{K}\right)$$

where $\alpha_{21}x_2$ and $\alpha_{12}x_1$ describe the constraint enforced by the growth of genotype-2 subpopulation on the growth of genotype-1 subpopulation, and vice versa, respectively. Generalized to a higher number of sub-populations, the change in representation of genotype ***i*** in the population at time interval ***t*** can be described as

$$\frac{dx_i}{dt} = \lambda_i x_i \left(1 - \frac{x_i + \sum_{j \neq i} \alpha_{ji} x_j}{K} \right)$$

which reduces back to the one-population case if $\alpha_{ij} = 1$ for all *i*,*j* pairs

We propagate individuals between consecutive generations (t-1) to (t) in two stages. First, a population (whose size can be different from that of the population at generation t-1) is formed in which the *i*–th genotype population size is given by its size in the previous generation and its fitness by:

$$x_{i(t)} \approx x_{i(t-1)} e^{\lambda_i}$$

Then, to keep a constant population size stochastic rescaling is applied that implements a Kimura-governed (5) allele sampling.

Calculation of tRNAs-to-codons ratio

We performed a rough analysis that aimed to assess the relative abundance of tRNAs and codons in the cell. In particular, we examined the six rarest tRNAs in *S. cerevisiae*, each of which is encoded in the yeast genome by only one tRNA gene. These six rare tRNAs correspond to seven codons: CGG (Arg), CAG (Gln), ACG (Thr), UCG (Ser), AGG (Arg), CUU (Leu) and CUC (Leu). There is one-to-one correspondence between each of the first four codons and their tRNA; Codon AGG can be also translated by the fully-matched tRNA of AGA (6); the last two codons are translated by the same tRNA type, hence are counted together.

Estimates suggest that a yeast cell contains some 3.3 million tRNA molecules (BioNumbers database (7) and (8)). The copy number of molecules of each tRNA type is simply the fraction of its tRNA gene copy number out of the total gene copy number of all tRNA types multiplied by 3.3 million. As for codons, the number of codons of any type in the transcriptome is defined by the sum of appearances of a codon along all genes in the genome, multiplied by the average mRNA abundance in the cell (9)). To consider specifically the subset of codons that are actively translated, we consider the fraction of mRNAs which are occupied by at least one ribosome (=0.71, (10)).

The table below shows the ratio of the number of tRNA molecules to the corresponding codon copy number for the above selection of codons. As can be seen, the ratio is never larger or smaller than 10, suggesting that tRNA and their respective codons are estimated to be in similar amounts in the cell.

Codon	tRNA/codon abundence
Arg (agg)	0.22
Arg (cgg)	1.15
Gln (cag)	0.17
Ser (ucg)	0.24
Thr (acg)	0.26
Leu (cuc & cuu)	0.12

References

1. Kudla, G., Lipinski, L., Caffin, F., Helwak, A. and Zylicz, M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*, **4**, e180.
2. Mazel, D. and Marliere, P. (1989) Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature*, **341**, 245-248.
3. Drake, J.W., Charlesworth, B., Charlesworth, D. and Crow, J.F. (1998) Rates of spontaneous mutation. *Genetics*, **148**, 1667-1686.
4. Joseph, S.B. and Hall, D.W. (2004) Spontaneous mutations in diploid Saccharomyces cerevisiae: more beneficial than expected. *Genetics*, **168**, 1817-1825.
5. J.F.Crow and Kimura, M. (1970) *An Introduction to Population Genetics Theory*. Harper & Row, New York.
6. Johansson, M.J., Esberg, A., Huang, B., Bjork, G.R. and Bystrom, A.S. (2008) Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Mol Cell Biol*, **28**, 3301-3312.
7. Milo, R., Jorgensen, P., Moran, U., Weber, G. and Springer, M. (2010) BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res*, **38**, D750-753.
8. Waldron, C. and Lacroute, F. (1975) Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. *J Bacteriol*, **122**, 855-865.
9. Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717-728.
10. Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O. and Herschlag, D. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, **100**, 3889-3894.
11. Shalem, O., Dahan, O., Levo, M., Martinez, M.R., Furman, I., Segal, E. and Pilpel, Y. (2008) Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol*, **4**, 223.

Figure S1: The fold-changes in representation of amino acids and nucleotide types in the transcriptome upon stress. (a) "Amino acid-Expression" matrix for diverse stress types, normalized as in Figure 2a. Each cell denotes the fold-change in the representation of a given amino acid in the transcriptome at a given time point upon specific stress, compared to its representation at time point zero. The amino acid labels are followed by numbers in parentheses, indicating the sum of gene copy number of all their corresponding tRNAs. (b) "Nucleotide-Expression" matrix for diverse stress types, normalized as in Figure 2a. Each cell denotes the fold-changes in the representation of a given nucleotide in the transcriptome at discrete time point (minutes) upon a specific stress, compared to its representation at the corresponding time point zero. A one letter label (a,c,g and u) refers to the total nucleotide representation, whereas specific codon position labels indicate the usage of a given nucleotide at each of the three positions of the codon.

Figure S2: Correlation between the codons adaptiveness values and the change in their representation in the transcriptome under stress. We calculated the Pearson correlation coefficient between 61-long vectors denoting fold-changes in the codon usage of the transcriptome in different time points (minutes) of diverse environmental conditions and the 61 codons' tAI values. A consistent negative correlation between the codons adaptiveness values (W_i) and their representation in the transcriptome in stress can be seen. The most negative correlations among the different time points in each of the examined stress types vary between -0.52 (oxidative stress) and -0.73 (MMS). Other than the correlation value for the first time point of the oxidative stress, all the correlations were found to be significant, with p-values spanning a range of 2.45×10^{-11} to 4.76×10^{-2}. The recovery from both heat-shock and the KCL stresses, (labeled 'R'), obtained by transferring the cells from the respective stressful conditions to normal growth conditions, is accompanied by sharp increase of the measured correlations between the codons' adaptiveness value (W_i) and their representation in the transcriptome, towards significant positive values (KCL: Pearson Correlation = 0.7, p-values = 4.43×10^{-10}; heat-shock: Pearson correlation = 0.67, p-values = 3.14×10^{-9}). We detected a similar pattern of change in the direction of the correlation, though with relatively moderate slope, for the oxidative stress, probably as a result of spontaneous recovery from the stress (11).

Figure S1

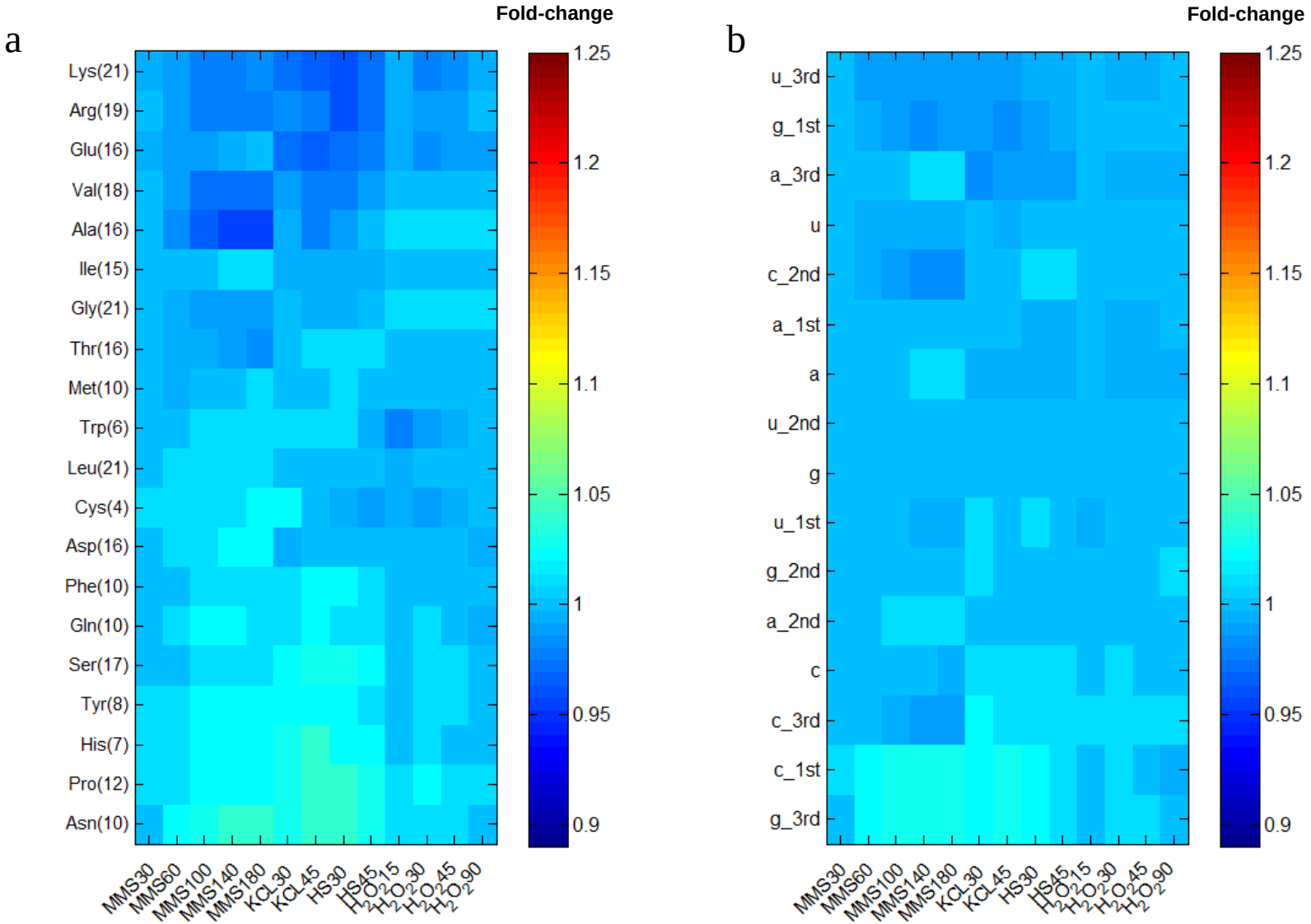
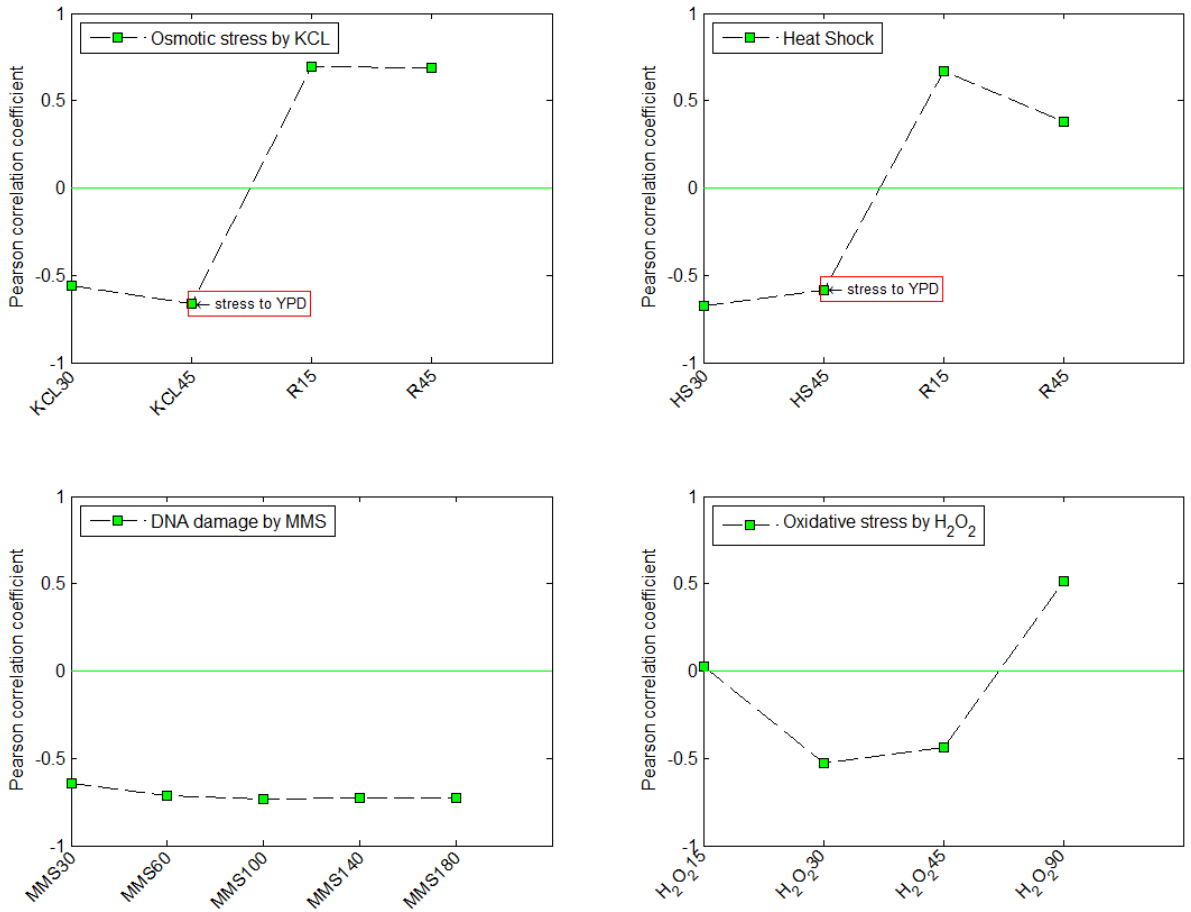

a
Fold-change
Lys(21)
Arg(19)
Glu(16)
Val(18)
Ala(16)
Ile(15)
Gly(21)
Thr(16)
Met(10)
Trp(6)
Leu(21)
Cys(4)
Asp(16)
Phe(10)
Gln(10)
Ser(17)
Tyr(8)
His(7)
Pro(12)
Asn(10)
MMS30
MMS60
MMS100
MMS140
MMS180
KCL30
KCL45
HS30
HS45
H2O215
H2O230
H2O245
H2O290
1.25
1.2
1.15
1.1
1.05
1
0.95
0.9
b
Fold-change
u_3rd
g_1st
a_3rd
u
c_2nd
a_1st
a
u_2nd
g
u_1st
g_2nd
a_2nd
c
c_3rd
c_1st
g_3rd
MMS30
MMS60
MMS100
MMS140
MMS180
KCL30
KCL45
HS30
HS45
H2O215
H2O230
H2O245
H2O290
1.25
1.2
1.15
1.1
1.05
1
0.95
0.9

Figure S2


Osmotic stress by KCL
Heat Shock
DNA damage by MMS
Oxidative stress by H_2O_2
Pearson correlation coefficient
stress to YPD
1
0.5
0
-0.5
-1
KCL30
KCL45
R15
R45
HS30
HS45
MMS30
MMS60
MMS100
MMS140
MMS180
H_2O_215
H_2O_230
H_2O_245
H_2O_290