

Evolutionary genomics

Codon bias and selection on single genomes

Arising from: J. B. Plotkin, J. Dushoff & H. B. Fraser *Nature* **428**, 942–945 (2004)

The idea that natural selection on genes might be detected using only a single genome has been put forward by Plotkin and colleagues¹, who present a method that they claim can detect selection without the need for comparative data and which, if correct, would confer greater power of analysis with less information. Here we argue that their method depends on assumptions that confound their conclusions and that, even if these assumptions were valid, the authors' inferences about adaptive natural selection are unjustified.

The volatility analysis of Plotkin *et al.*¹ rests on the observation that synonymous codons often differ in the number of mutations that take them to different amino acids. For instance, CGA and AGA both code for arginine, but differ in the number of amino acids that are one mutation away (4 out of 8 for CGA and 6 out of 8 for AGA; AGA therefore has a higher volatility). Their expectation is that proteins that have undergone more amino-acid substitutions will have more highly volatile codons.

To test each gene for selection by looking at relative codon volatility, Plotkin *et al.* construct the null by drawing alternative synonymous codons from a distribution parameterized by genome-wide codon frequencies, with a view to identifying genes that are atypically rich or poor in volatile codons and which they believe represent rapidly and slowly evolving genes, respectively.

A premise of this approach is that codon usage does not vary in a consistent way from gene to gene. If codon bias is related to volatility in any way — for instance, if CGA is preferentially used over AGA in highly expressed genes — then the volatility index that the authors use is simply an alternative measure of within-genome variation in codon-usage bias. In fact, differential codon bias due to both natural selection and mutation bias results in a highly heterogeneous distribution of codon usage across multiple genomes^{2–5}.

To investigate the extent to which the authors' unmet assumption will affect estimation of volatility, we conducted two analyses. First, we examined the correlation between volatility *P* values and the codon adaptation index (CAI), a common measure of optimal codon bias, for every gene in the *Saccharomyces cerevisiae* genome (Fig. 1a). We use *S. cerevisiae* because the optimal codons are known⁶ and because nucleotide divergence can be estimated from a closely related outgroup (*S. paradoxus*). As expected, we found a strong correlation between CAI and volatility, with CAI explaining a much larger proportion of the variance in volatility than the standard comparative measure of selection, dN/dS.

That codon usage bias — measured by either CAI or volatility — correlates with selective constraint is well known and unsurprising^{7,8}.

Second, because CAI measures only one of many known biases in codon usage, to examine the general effects of differential codon usage we randomly assigned volatility scores to each codon and then re-analysed the *Mycobacterium tuberculosis* genome according to the method of Plotkin *et al.*¹. Figure 1b shows the distribution of volatility scores for this randomized genome. This distribution has a U-shape similar to that found in the true volatility distribution for *M. tuberculosis*, and there is high similarity in the set of genes that reside in the tails of both distributions ($P < 10^{-15}$; Fig. 1b); these outliers show dissimilar codon usage from the majority of the genes, regardless of the measure used. That a random assignment of codon volatilities recovers the same outlier genes as the true values suggests that volatility itself has little to do with the observed distribution. Our analyses indicate that volatility may be another measure of codon bias.

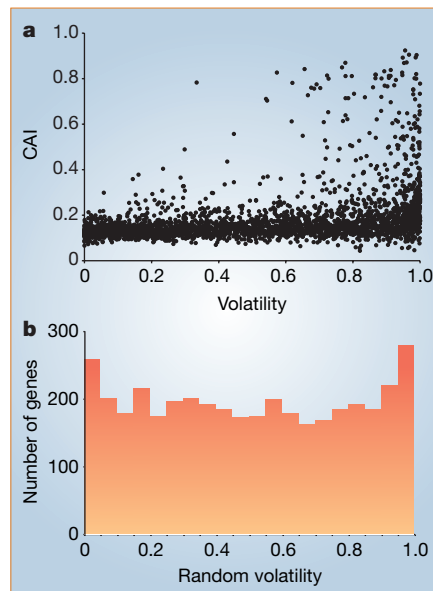


Figure 1 Codon bias explains volatility. **a**, Relationship between the codon adaptation index (CAI) and the volatility *P* values in the *Saccharomyces cerevisiae* genome. An analysis of variance with CAI and dN/dS (calculated from a comparison with orthologs in *S. paradoxus*) as main effects reveals that CAI (*F* ratio = 322, $P < 0.0001$) explains more than twice as much of the variation in volatility as dN/dS (*F* ratio = 153, $P < 0.0001$). **b**, Distribution of random volatilities in the *Mycobacterium tuberculosis* genome. We randomly assigned each codon a volatility score between 0 and 1 and calculated the volatility *P* values for all the genes in the *M. tuberculosis* genome using these new values, following the method of Plotkin *et al.*¹. The proportion of the 78 genes in the 1% tails of the volatility distribution that are shared between the random and true volatility distributions is 35% (27 genes in both). This overlap is highly non-random ($P < 10^{-15}$).

The authors do not provide formal reasoning to explain why the volatility statistic should correlate with selective constraint. If their method is grounded in standard, single-locus multiple-allele models, their result is trivial: in such models, assigning an allele lower fitness will deterministically lower its frequency in a population. And because volatility only applies to four codon families, three of which contain synonymous codons that cannot be reached by a single mutation, any such population-genetics model involving volatility must violate basic assumptions.

The authors claim both that their method does not rely on some of the strongest assumptions of comparative analyses¹ and that their unstated model assumes a population at mutation–selection balance. However, comparative methods for detecting natural selection (see ref. 9, for example) do not require populations in mutation–selection balance. Codon-based comparative methods do require a mutational process at stationarity⁹, but so do Plotkin *et al.*: if they do not assume mutational stationarity, their logic does not work.

Consider pseudogenes: volatility scores of these genes will reflect past processes, not current selection, until they reach a stationary state. Volatility is only expected to be associated with amino-acid turnover as a genome approaches codon-usage stationarity, but it is not clear how to test the assumption that a genome has reached such a state.

The distribution of negative selection across a genome is unknown and, because of this — even if all the assumptions of the model have been met — one cannot say with certainty that genes in the tails of the distribution are under increased positive selection or decreased negative selection. Simply because there are a handful of genes thought to be under positive selection present in these tails does not prove that all or even most of these genes are. Although Plotkin *et al.* provide a caveat to inferences about positive selection in their penultimate paragraph¹, this is undermined by the seven preceding claims of positive selection in the paper. Comparative analyses such as dN/dS do require data from multiple taxa, but they provide a clear statistical criterion for detecting positive selection.

Matthew W. Hahn, Jason G. Mezey, David J. Begun, John H. Gillespie, Andrew D. Kern, Charles H. Langley, Leonie C. Moyle

Center for Population Biology and Section of Evolution and Ecology, University of California, Davis, California 95616, USA

e-mail: mwhahn@ucdavis.edu

doi:10.1038/nature03221

- Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428**, 942–945 (2004).
- Akashi, H. *Genetics* **136**, 927–935 (1994).
- Chiappello, H. *et al. Gene* **209**, GC1–GC38 (1998).
- Coghlan, A. & Wolfe, K. H. *Yeast* **16**, 1131–1145 (2000).
- Marais, G., Mouchiroud, D. & Duret, L. *Proc. Natl. Acad. Sci. USA* **98**, 5688–5692 (2001).
- Ikemura, T. *Mol. Biol. Evol.* **2**, 13–34 (1985).

7. Sharp, P. M. *J. Mol. Evol.* **33**, 23–33 (1991).
 8. Akashi, H. *Curr. Opin. Gen. Dev.* **11**, 660–666 (2001).
 9. Goldman, N. & Yang, Z. *Mol. Biol. Evol.* **11**, 725–736 (1994).
 Reply: J. B. Plotkin, J. Dushoff and H. B. Fraser reply to this communication (doi:10.1038/nature03224).

Evolutionary genomics

Detecting selection needs comparative data

Positive selection at the molecular level is usually indicated by an increase in the ratio of non-synonymous to synonymous substitutions (dN/dS) in comparative data. However, Plotkin *et al.*¹ describe a new method for detecting positive selection based on a single nucleotide sequence. We show here that this method is particularly sensitive to assumptions regarding the underlying mutational processes and does not provide a reliable way to identify positive selection.

Plotkin *et al.*¹ use a measure for detecting selection known as the volatility index, whereby a codon with high volatility is more likely to have arisen by a non-synonymous mutation than a codon with low volatility; so, for high dN/dS , there should be more codons of high volatility. Positive selection should be detectable simply by examining the volatility index in a single sequence.

However, this argument is flawed because high rates of non-synonymous mutation will increase the rate of substitution both into and out of codons with high volatility. In models in which the substitution process is reversible over time, these two factors will cancel each other out, and variations in the strength of selection at the amino-acid level do not affect the expected volatility. Although most models used in studies of molecular evolution are time-reversible², the true substitution process probably is not, because of the specifics of the mutational and population-level processes.

To examine the effect of the substitution model on the volatility index, we simulated random-substitution models in which the rate of substitution between different nucleotides was sampled from a uniform random variable between zero and one. For these models, we then calculated the equilibrium frequencies of the 61 sense codons in a Markov chain model that resulted from simulations having varying synonymous and non-synonymous substitution rates. Based on the equilibrium frequencies, we could then calculate the expected value of the volatility index.

Our results indicate that the volatility index can be either an increasing or a decreasing function of dN/dS , or have a minimum or maximum at an intermediate value of this ratio (Fig. 1). We also find that the dN/dS ratio only marginally affects the volatility index — particularly for values of

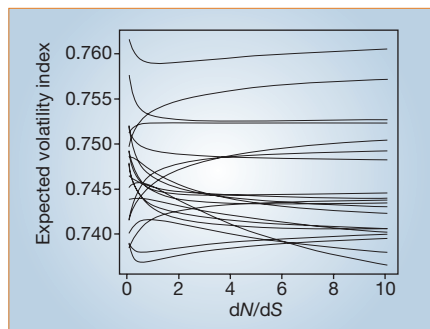


Figure 1 Expected value of the volatility index, as defined by Plotkin *et al.*¹, as a function of the dN/dS ratio for 20 random-substitution models.

$dN/dS > 1$. Although models can be constructed in which strong stabilizing selection on particular amino acids has a marked effect on the volatility index, there is no evidence that the volatility index captures much information regarding positive selection. Realistic models of positive selection will predict an increased rate of substitution both in and out of codons with high volatility.

What then explains the results of Plotkin *et al.*¹, in which the volatility index correlates with the rate of amino-acid substitution in comparative data and with the amount of expression? Non-random codon usage is common in most organisms, particularly in bacteria and yeast^{3–7}, and may be caused by selection for optimal codon usage and affected by variation in the nucleotide composition and other factors. In bacteria, the strength of codon-usage bias is correlated with the amount of expression^{3–5} and with the extent of amino-acid substitution^{6,7}; this may be because highly expressed genes tend to be more conserved at the amino-acid level and have more codon-usage bias than genes with low expression. The degree of amino-acid substitution might also correlate with local nucleotide frequencies because regions that differ in this respect could have different rates and patterns of mutation.

To investigate the extent to which the volatility index is sensitive to local nucleotide content, we took advantage of the fact that only codons with sixfold degeneracy or with stop codons as neighbours can contribute to the volatility index. Using all other codons we obtained independent estimates of the nucleotide frequencies. We also calculated a P value for a one-tailed test of increase in the frequency of a particular nucleotide by using the methodology of Plotkin *et al.*¹, but calculated only for codons that do not contribute to the volatility index.

Applying this approach to the *Plasmodium falciparum* data analysed by Plotkin *et al.*¹, the correlation coefficient between the log P value of the volatility index and the log P value associated with the percentage of thymine is 0.29. Variation in third-position nucleotide content is one of the factors explaining the distribution of volatility-

index-related P values in *P. falciparum*. Correlation of the volatility index with the amount of amino-acid substitution could be caused by the presence of covariates such as nucleotide frequencies, selection for optimal codon usage bias and/or expression levels.

The results of Plotkin *et al.*¹ might also be explained by variation in the amino-acid frequencies among genes. If the true evolutionary model is not time-reversible, these frequencies should influence codon usage and the volatility P value. Indeed, many of the amino-acid frequencies show correlation with the volatility P values calculated by Plotkin *et al.*¹. For example, the correlation coefficient between the frequency of glutamine and the log volatility P values is -0.32 . All codons for glutamine have the same volatility, but this amino acid is one mutational step away from arginine and leucine, which both affect the volatility index. The volatility index in models that are not time-reversible can therefore be affected by stabilizing selection on particular amino acids, because such selection affects the amino-acid frequency. But whether the volatility index correlates positively or negatively with such selection depends on which amino acid is the target of selection. Positive selection that increases the rate of amino-acid substitution does not have the same impact on the volatility index.

We argue that the volatility index cannot be applied to detect positive selection as it is under greater influence from other factors, such as amino-acid and nucleotide frequencies. However, the results of Plotkin *et al.*¹ should spur efforts to identify the causes of non-random codon usage in bacteria and other organisms.

Rasmus Nielsen*†, Melissa J. Hubisz†

*Centre for Bioinformatics, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark
 e-mail: rasmus@binf.ku.dk

†Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA

doi:10.1038/nature03222

- Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428**, 942–945 (2004).
- Lio, P. & Goldman, N. *Genome Res.* **8**, 1233–1244 (1998).
- Ikemura, T. *J. Mol. Biol.* **151**, 389–409 (1981).
- Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. *Nucl. Acids Res.* **9**, 43–74 (1981).
- Grosjean, H. & Fiers, W. *Gene* **18**, 199–209 (1982).
- Sharp, P. M. *J. Mol. Evol.* **33**, 23–33 (1991).
- Akashi, H. & Gojobori, T. *Proc. Natl Acad. Sci. USA* **99**, 3695–3670 (2002).

Reply: J. B. Plotkin, J. Dushoff and H. B. Fraser reply to this communication (doi:10.1038/nature03224).

Evolutionary genomics

Codon volatility does not detect selection

Plotkin *et al.*¹ introduce a method to detect selection that is based on an index called codon volatility and that