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 (d*N*/d*S*) 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 d*N*/d*S*, 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-reversible[2], 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 d*N*/d*S*, or have a minimum or maximum at an intermediate value of this ratio (Fig. 1). We also find that the d*N*/d*S* ratio only marginally affects the volatility index — particularly for values of d*N*/d*S* > 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.

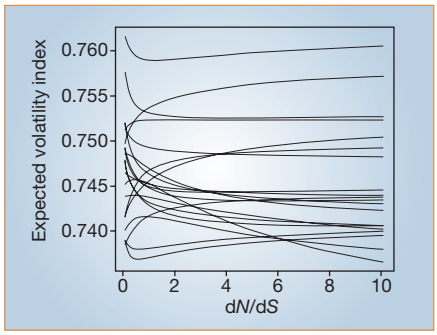

Figure 1 Expected value of the volatility index, as defined by Plotkin *et al.*[1], as a function of the d*N*/d*S* ratio for 20 random-substitution models.

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

1. Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428,** 942–945 (2004).
2. Lio, P. & Goldman, N. *Genome Res.* **8,** 1233–1244 (1998).
3. Ikemura, T. *J. Mol. Biol.* **151,** 389–409 (1981).
4. Grantham, R., Gautier, C., Gouy, M., Jacobzone., M & Mercier, R. *Nucl. Acids Res.* **9,** 43–74 (1981).
5. Grosjean, H. & Fiers, W. *Gene* **18,** 199–209 (1982).
6. Sharp, P. M. *J. Mol. Evol.* **33,** 23–33 (1991).
7. 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

uses only the sequence of a single genome, claiming that this method is applicable to a large range of sequenced organisms[1]. Volatility for a given codon is the ratio of non-synonymous codons to all sense codons accessible by one point mutation. The significance of each gene's volatility is assessed by comparison with a simulated distribution of 10^6 synonymous versions of each gene, with synonymous codons drawn randomly from average genome frequencies. Here we re-examine their method and data and find that codon volatility does not detect selection, and that, even if it did, the genomes of *Mycobacterium tuberculosis* and *Plasmodium falciparum*, as well as those of most sequenced organisms, do not meet the assumptions necessary for application of their method.

Because of codon table regularity, 16 of 20 codon families have the same volatility for each synonymous codon. The authors' assertion that there are 22 codons that have at least one synonym with a different volatility[1] obscures the fact that these 22 codons represent only four amino acids: the four-fold-degenerate glycine codons and the six-fold-degenerate arginine, leucine and serine codons. Therefore, even if the method were valid, volatility neglects any selection acting on the remaining codons. Figure 1 shows that the volatility *P* values of Plotkin *et al.*[1], which are purported to detect selection, are largely a measure of the codon usage of serine, with smaller contributions from glycine, arginine and leucine. A codon-usage index dominated by a single family is not generally useful for studying codon bias, much less selection — unless the genic signature of natural selection is merely serine-codon usage.

The theoretical basis of the method of Plotkin *et al.*[1] is unclear: either it depends critically on the only theoretical justification offered by the authors, a model by van Nimwegen *et al.*[2], or it is unfounded. Assuming the null model is that of Nimwegen *et al.*, it requires that all synonymous codons are of equal fitness, that all members of a family are mutually accessible by a path of neutral mutation, and that the product of the effective population size and mutation rate, $N_e\mu$, is much greater than one. But these requirements are respectively violated in that selection for codon usage in highly expressed genes is common[3]; the twofold-degenerate serine codons cannot access the fourfold-degenerate codons without passing through a non-synonymous intermediate; and most sequenced genomes have an estimated $N_e\mu \ll 1$ (ref. 4). Specifically, $N_e\mu$ for pathogens is constrained by repeated bottlenecks caused by transmission. In fact, the highest estimated $N_e\mu$ values for both *P. falciparum* and *M. tuberculosis* are of the order of 10^{-3} or less[5,6].

In Table 1 of Plotkin *et al.*[1], the authors present ten genes (including five of putative function, not identified as such, and one hypothetical protein) that they say show the most significant volatility *P* values and that therefore "show the strongest signs of positive selection", although they ignore 27 other genes (see their supplementary information) with *P* values indistinguishable from those in their table. These include a putative cell-cycle control protein, histone deacetylase, and DNA-directed RNA polymerase — highly conserved genes not expected to show signs of recent positive selection. Furthermore, genes encoding transfer RNA and ribosomal RNA are included in their analysis and assigned volatility *P* values, although these genes do not have codons.

The reality at present is that the community must continue to rely on other methods to detect natural selection.

Ying Chen*, J. J. Emerson*, Todd M. Martin†
**Department of Ecology and Evolution, and †Committee on Genetics, University of Chicago, Chicago, Illinois 60637, USA*

doi:10.1038/nature03223

1. Plotkin, J. B., Dushoff, J. & Fraser, H. B. *Nature* **428,** 942–945 (2004).
2. van Nimwegen, E., Crutchfield, J. P. & Huynen, M. *Proc. Natl Acad. Sci. USA* **96,** 9716–9720 (1999).
3. Akashi, H. *Curr. Opin. Genet. Dev.* **11,** 660–666 (2001).
4. Lynch, M. & Conery, J. S. *Science* **302,** 1401–1404 (2003).
5. Hughes, A. L. & Verra, F. *Proc. R. Soc. Lond.* **268,** 1855–1860 (2001).
6. Sreevatsan, S. *et al. Proc. Natl Acad. Sci. USA* **94,** 9869–9874 (1997).

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

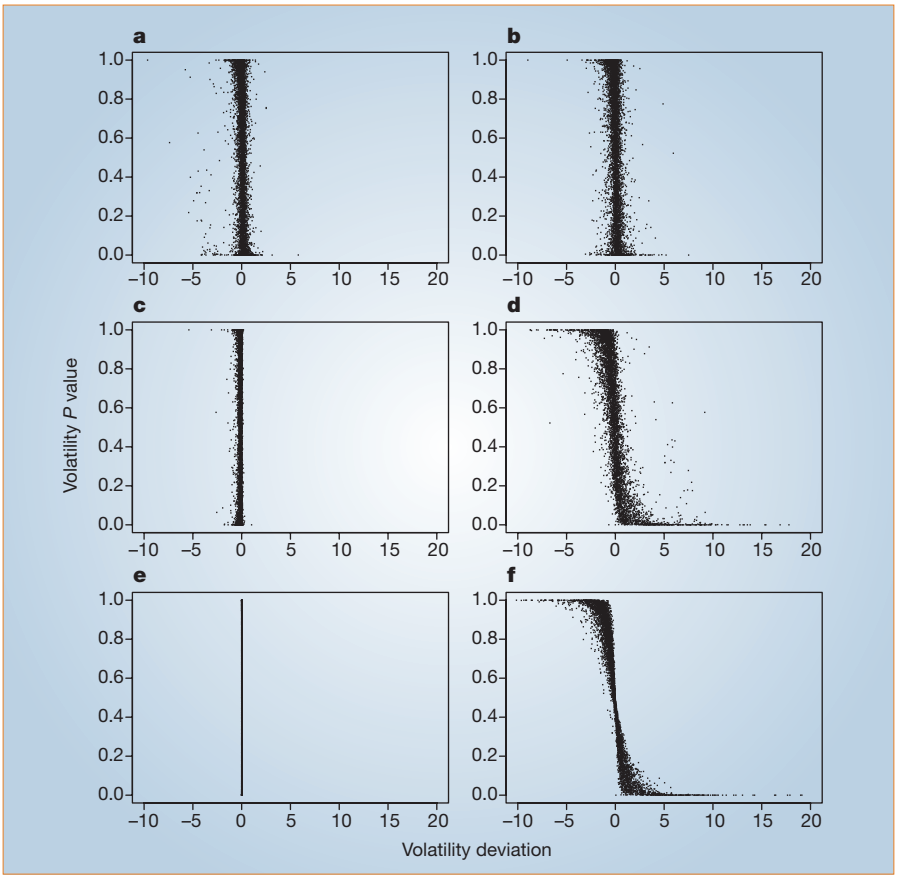

Figure 1 Volatility *P* value plotted against volatility deviation. For each gene in the *Plasmodium falciparum* genome, the relationship between the volatility *P* values of Plotkin *et al.*[1], which reportedly consider all codon contributions, and the difference between the observed volatility contributions for given codons and the volatility contributions expected by their genome average frequency is shown. Serine is the only amino acid for which the tails of the statistic are significantly associated with extremes in the *P* value. This analysis was repeated for each gene in the *Mycobacterium tuberculosis* genome (data not shown). **a,** Arginine; **b,** leucine; **c,** glycine; **d,** serine; **e,** all other families, excluding arginine, leucine, glycine and serine; and **f,** all 20 families.

Plotkin et al. *reply* — The criticisms of our results[1] by Hahn *et al.*[2], Nielsen and Hubisz[3], and Chen, Emerson and Martin[4] fall into three categories: the formal justification for our method, the potential for confounding factors, and the interpretation of our empirical results.

In response to assertions[2,4] that we do not formally discuss why codon volatility should reflect selection pressures on proteins, our method is grounded in the standard multi-allele model of population genetics[5,6]. Hahn *et al.* note that assigning an allele lower fitness will deterministically lower its frequency in a population[2], but this fundamentally misunderstands why volatility reflects selection: even if all synonymous codons for an amino acid are assigned equal fitness, selection at the amino-acid level will bias the relative frequencies of synonymous codons.

Consider, for example, a site under negative selection for arginine. Assigning equal fitness to arginine codons and lower fitness to all other codons, the multi-allele model[5] indicates that a less volatile arginine