

Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*

Joshua B. Plotkin, Jonathan Dushoff, Hunter B. Fraser

Supplementary Methods

Volatility versus comparative sequence analysis

For comparison with volatility, we have analyzed the complete genomes of three organisms related to *M. tuberculosis* strain CDC1551: *M. tuberculosis* strain H37Rv, *M. bovis*, and *M. leprae*. We have compared volatility p -values in *M. tuberculosis* against dN. For large-scale pairwise analysis, dN is often a more stable indicator of recent amino acid changes than dN/dS, because dS can be strongly influenced by selection on synonymous sites due to codon bias^{1, 2}. We identified putative orthologs using the method of reciprocal smallest distances, which is more reliable than reciprocal best BLAST hits³. Of the 4099 CDC1551 genes, 3502 orthologs were identified in H37Rv, 3479 orthologs in *M. bovis*, and 1354 orthologs in *M. leprae*.

The difficulty in finding orthologs is a systematic limitation of all comparative methods. Moreover, often the genes of greatest interest – *i.e.* those under selection for amino acid change – will diverge more quickly, further reducing the chance of identifying orthologs. Indeed, we note that the set of *M. tuberculosis* genes with identifiable orthologs in *M. leprae* is significantly less volatile than the genes without orthologs ($p_w = 10^{-6}$). For example, only six of the 57 PPE genes in *M. tuberculosis* have identifiable orthologs in *M. leprae*. (*P. falciparum* and its closest fully sequenced relative, *P. yoelii yoelii*, have only 2432 identifiable orthologs, none of which includes var, rifin, or stevor.)

We aligned the protein sequences of putative orthologs using the package ClustalW⁴. After alignment, we estimated dN using the maximum likelihood package PAML⁵. Maximum likelihood estimation assumes that substitutions follow a time-homogeneous Markov process, and that substitution rates vary across sites according to a fixed distribution⁵. We note that the method of differential codon volatility makes neither of these assumptions.

We have already seen that dN values and volatility p -values detect positive selection in the same gene families of *M. tuberculosis*⁶. On a gene-by-gene basis, we find a highly significant, but weak, correlation between volatility and dN (CDC1551 volatility p -values versus dN to H37Rv,

Spearman $r = -0.16$, $p = 5 \cdot 10^{-9}$; versus dN to *M. bovis*, $r = -0.14$, $p = 2 \cdot 10^{-6}$; versus dN to *M. leprae*, $r = -0.12$, $p = 2 \cdot 10^{-5}$) for those genes that show at least one amino acid difference between species. These correlations demonstrate that differential codon volatility is consistent with comparative methods. It may not be reasonable to expect stronger correlations, considering the multitude of assumptions made in each stage of the genome-wide comparisons, and the lack of accuracy when comparing only two sequences. When extensive comparative data are available, there is a stronger correlation between volatility and dN or dN/dS: the volatility p -values for 3031 *S. cerevisiae* genes with identified orthologs in *S. bayanus*, *S. paradoxus*, and *S. mikatae* exhibit a Spearman correlation of $r = -0.38$ ($p = 10^{-124}$) compared against dN/dS estimated using all four yeast species (J.B.P. & H.B.F., manuscript in preparation).

Calculation of volatility p -values

We define the volatility of a gene G as the summed volatility of its codons. In order to compute the volatility p -value for gene G , we employ a statistical test that evaluates whether or not $\nu(G)$ is significantly elevated or depressed compared to the rest of the genome, controlling for the length and amino acid composition of G . In the main text, we describe a randomization procedure to compute the volatility p -value of each gene. We supplement that procedure here by describing a more direct calculation of the volatility p -value, based upon a normal approximation, which agrees with the randomization procedure used in the main text.

We index the 61 sense (*i.e.* non-stop) codons in an arbitrary order $i = 1 \dots 61$. We use the notation $\text{aa}(i)$ to denote the amino acid encoded by codon i . Let N_i and n_i denote the number of occurrences of codon i in the entire genome and in the gene G , respectively. Similarly let M_a and m_a denote the number of occurrences of amino acid a in the entire genome and in the gene G . For each amino acid, a , we define its expected volatility and its variance in volatility, given the codon usage in the entire genome, by the equations:

$$\begin{aligned} E[\nu(a)] &= \sum_{\text{aa}(i)=a} \nu(i) N_i / M_{\text{aa}(i)} \\ V[\nu(a)] &= \sum_{\text{aa}(i)=a} \nu(i)^2 N_i / M_{\text{aa}(i)} - E[\nu(a)]^2 \end{aligned}$$

Similarly, we define the expected volatility of gene G and its variance by the equations:

$$\begin{aligned} E[\nu(G)] &= \sum_{a=1}^{20} m_a E[\nu(a)] \\ V[\nu(G)] &= \sum_{a=1}^{20} m_a V[\nu(a)] \end{aligned}$$

We calculate the volatility p -value for G by comparing the gene’s observed volatility to its expected volatility, given the amino acid content of the gene and the codon usage in the entire genome. Since genes typically contain at least 30 amino acids, we use a normal approximation to calculate the volatility p -value:

$$p = \frac{1}{2} \left(1 + \text{Erf} \left(\frac{E[\nu(G)] - \nu(G)}{\sqrt{2V[\nu(G)]}} \right) \right)$$

where Erf denotes the error function,

$$\text{Erf}(x) = \frac{2}{\pi} \int_0^x e^{-t^2} dt.$$

Volatility p -values caculated using this normal approximation are virtually identical to p -values calculated via the randomization procedure ($r^2 > 0.99$ for each genome).

References

1. Sharp, P. M. and Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38 (1986).
2. Hirsh, A. E., Fraser, H. B., and Wall, D. P. Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Genetics*, submitted (2003).
3. Wall, D. P., Fraser, H. B., and Hirsh, A. E. Detecting putative orthologs. *Bioinformatics* **19**, 1710–1711 (2003).
4. Thompson, J. D., Higgins, D. G., and Gibson, T. J. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
5. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**, 555–556 (1997).
6. Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., De-Boy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J. F., Nelson, W. C., Umayam, L. A., Ermolaeva, M., Salzberg, S. L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs, W. R., Venter, J. C., and Fraser, C. M. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**, 5479–5490 (2002).