Supporting Online Material for

September 24, 2009

Estimates of the effect of natural selection on protein coding content

V. B. Yap*, H. Lindsay, S. Easteal, G. A. Huttley*

* — to whom correspondence should be addressed: Gavin.Huttley@anu.edu.au or stayapvb@nus.edu.sg

S1. Neutral process varies across mammal genome

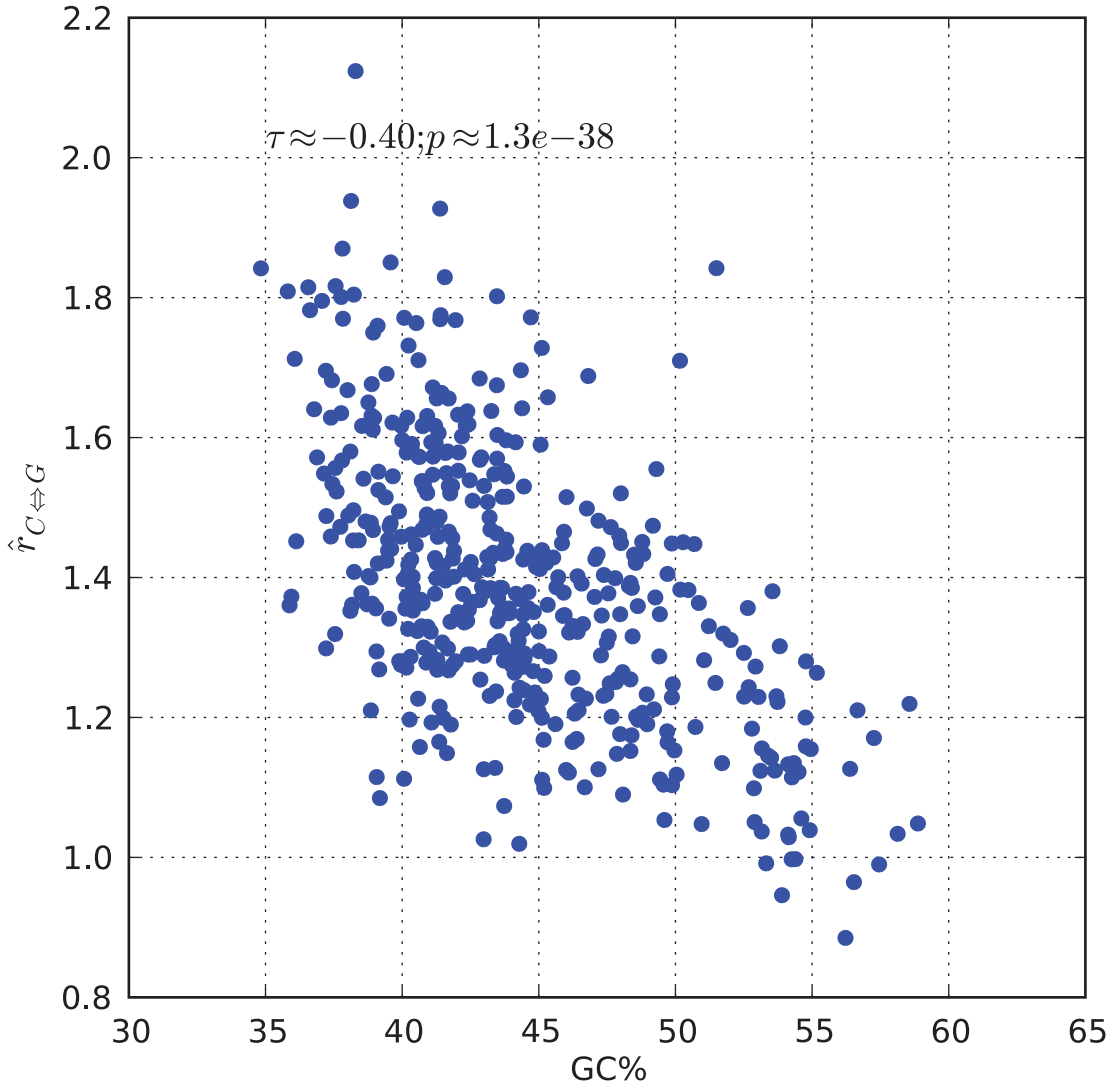
Data were alignments of intron sequences from 470 orthologous human, chimpanzee and macaque protein coding genes obtained from Ensembl release 50. We analysed the alignments with the nucleotide GTR model. In Fig. S1 we demonstrate a significant relationship between maximum likelihood estimates of the $r_{C\leftrightarrow G}$ parameter with alignment GC%. The correlation between $\hat{r}_{C\leftrightarrow G}$ and GC% was tested using Kendall's τ.

S2. Estimates of ω from primate introns vary with composition

Real mutation processes are much more complex than those used in our simulation study. For instance, evidence from vertebrates indicates a substantial proportion of mutations depend upon flanking nucleotides [2, 4, 7], on transcription [6] and local recombination rate [1, 5]. Many of these mutagenic influences exhibit spatial heterogeneity within genomes [4], and this variation in neutral processes has been associated with the striking compositional differences between genomic regions [3]. As shown above, this variation in neutral processes does affect nucleotide substitution model parameter estimates (Fig. S1). Because of this complexity, we evaluated whether the models were robust to evolutionary forces operating on real sequences that do not encode proteins.

We chose primate intronic sequences for assessment of model robustness to the complexity of neutral evolutionary processes for three reasons. First, introns have a mutation environment closely matching their flanking exons. The complex array of mutagenic influences affect both exons and their adjacent intronic sequences. These include context-dependent substitution processes which result in trinucleotide frequencies being significantly non-multiplicative. They also include the influence of biased gene conversion, which has been argued to contribute to overestimation of ω [1, 5]. Second, intron sequence evolution is not affected by selective constraint for codon usage or protein coding content. Third, the proportion of intronic sequences known to be functional is relatively modest [8]. Data were the same alignments of intron sequences described above. The intron alignments were treated as if they were in-frame protein coding sequences. Sequence regions

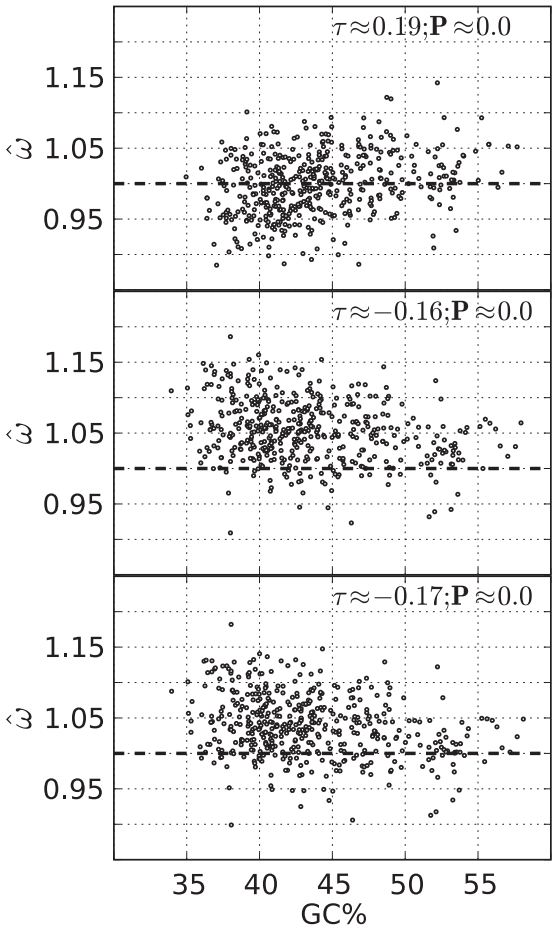
Figure S1 An example nucleotide substitution rate parameter that varies by composition. Data were primate intron alignments. GC% – the mean G+C nucleotide percentage from an alignment; $\hat{r}_{C \leftrightarrow G}$ – MLE of the relative rate of C/G transversions estimated from an alignment under the nucleotide GTR model.



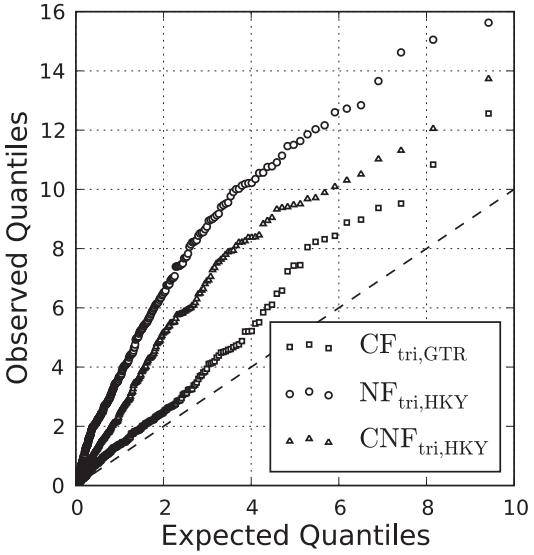
likely to evolve in a non-point mutation like manner, including indels, were removed. Results from analysis of these alignments, which evolve in a 'neutral' manner with respect to their hypothetical protein coding content, with the $CF_{tri,GTR}$, $NF_{tri,HKY}$ and $CNF_{tri,HKY}$ forms are shown in Fig. S2a. These models showed stronger sensitivity of $\hat{\omega}$ to GC% than those presented in the main manuscript. We further considered the consistency of the LR statistic quantiles with the expected quantile distribution under the null ($\omega = 1$). All models exhibited striking departures from the expected distribution Fig. S2b.

Figure S2 The model variants for which $\hat{\omega}$ was most biased by GC%. Data were from primate intron alignments.

(a) Scatter of $\hat{\omega}$ vs GC%



(b) Quantile-quantile plot



S3. Sensitivity of ω_{NF} to non-multiplicative codon frequencies

The NF model form has equilibrium frequency vectors that are multiplicative. Departures from this condition are then expected to cause parameter estimates under NF to be biased. This is demonstrated in the main manuscript for simulated data. We now evaluate whether this effect is evident in real biological sequences by comparing a measurement of the non-multiplicative nature of codon frequencies against a measure of their discordance in estimates of ω. We were specifically interested in establishing the consequence of failing to account for non-multiplicative π and not the difference between CNF and NF due to

their handling of stop codons. We therefore compared the $\text{CNF}_\times$ model, rather than NF, with CNF. This comparison ensures differences in $\hat{\omega}$ arise solely from specification of π.

We used a LR as a measure of the discordance between the MLE of ω from the multiplicative ($\hat{\omega}_{\text{CN}_\times}$) and non-multiplicative ($\hat{\omega}_{\text{CN}}$) CNF model forms. The $\hat{\omega}_{\text{CN}}$ was obtained by maximising the likelihood of a CNF_{GTR} model where the codon frequencies were estimated directly from the alignment. The $\hat{\omega}_{\text{CN}_\times}$ was obtained by maximising the likelihood of a $\text{CNF}_{\times,\text{GTR}}$ model where the codon frequencies were calculated from the alignment nucleotide frequencies. Discordance was measured under the CNF_{GTR} model. Specifically, we constrained $\omega_{\text{CN}} = \hat{\omega}_{\text{CN}_\times}$ and maximised the likelihood. The standard LR statistic was then computed between this constrained model and the unconstrained (free) CNF model as $\text{LR}= 2[\ln L(\text{free } \omega_{\text{CN}}) - \ln L(\omega_{\text{CN}} = \hat{\omega}_{\text{CN}_\times})]$. The resulting LR was employed as a measure of the discordance in ω arising from assuming multiplicative π. We used the χ^2 statistic from the standard goodness-of-fit test as a measure of the departure from multiplicative codon frequencies. Scatter plots between the statistics measuring discordance and non-multiplicative codon frequencies are shown in Fig. S3. The association between these statistics was tested using Kendall's τ.

For comparability, we computed discordance between ω estimated under the CF form with that estimated under CNF in an analogous manner to that described above. In this case, the multiplicative $\text{CNF}_{\times,\text{GTR}}$ model was replaced by CF_{HKY} and $\omega_{\text{CN}_\times}$ by ω_{CF}. The results of this analysis are shown in Fig. S4.

References

[1] J. Berglund, K. S. Pollard, M. T. Webster. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol*, 7(1):e26, Jan 2009.

[2] D. N. Cooper, E. V. Ball, M. Krawczak. The human gene mutation database. *Nucleic Acids Res*, 26(1):285–287, Jan 1998.

[3] L. Duret, A. Eyre-Walker, N. Galtier. A new perspective on isochore evolution. *Gene*, 385:71–74, Dec 2006.

[4] N. Elango, S.-H. Kim, E. Vigoda, S. V. Yi. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. *PLoS Comput Biol*, 4(2):e1000015, Feb 2008.

[5] N. Galtier, L. Duret, S. Glemin, V. Ranwez. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*, 25(1):1–5, Jan 2009.

[6] P. Green, B. Ewing, W. Miller, P. J. Thomas, E. D. Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*, 33(4):514–7, 2003.

[7] H. Lindsay, V. B. Yap, H. Ying, G. A. Huttley. Pitfalls of the most commonly used models of context dependent substitution. *Biol Direct*, 3:52, 2008.

[8] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

Figure S3 Sensitivity of ω to non-multiplicative codon frequencies. The $\mathrm{LR}(\omega_{\mathrm{CNF}} = \hat{\omega}_{\mathrm{CNF}_\times})$ measures the effect of not-accounting for non-multiplicative codon frequencies. χ^2(multiplicative) measures the extent to which observed codon frequencies were non-multiplicative. The calculation of the statistics is described in the text.

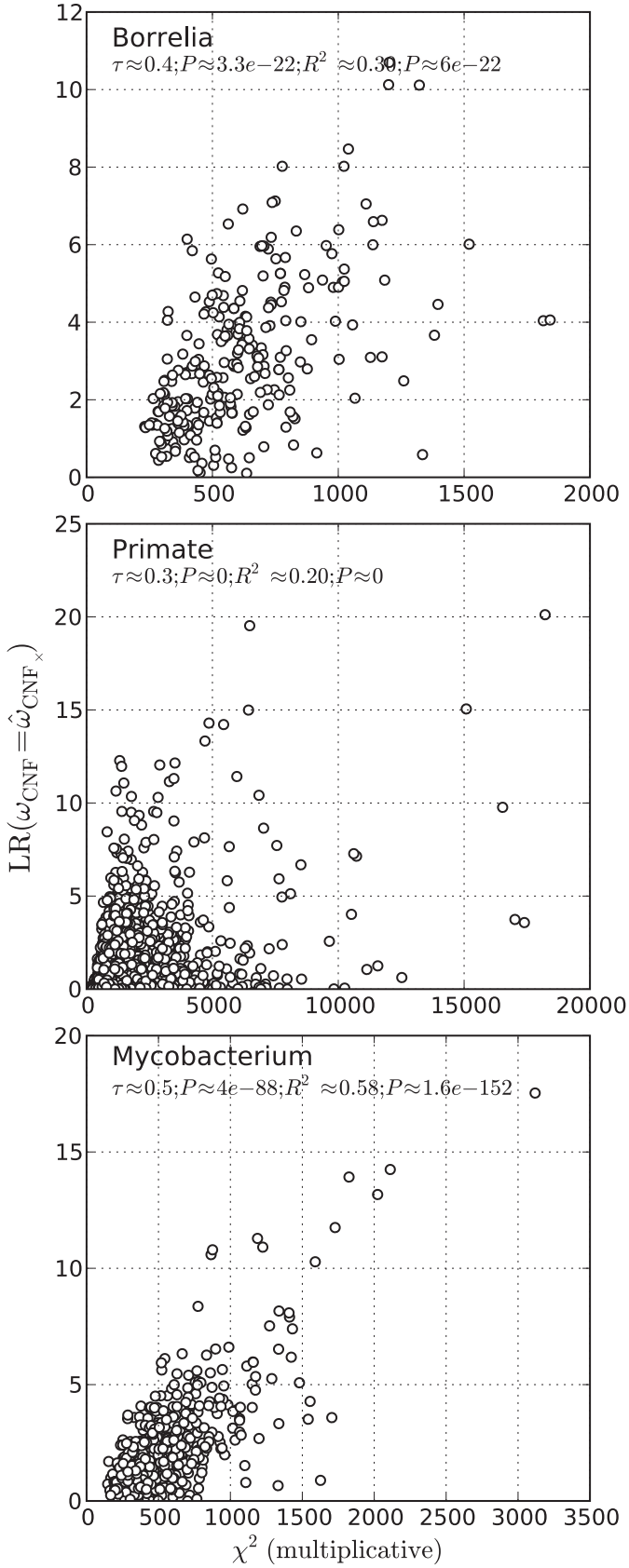


Figure S4 Discordance of $\hat{\omega}_{\mathrm{CNF}}$ and $\hat{\omega}_{\mathrm{CF}}$. We modified the discordance measure described in Fig. S3, replacing the $\mathrm{CNF}_{\times,\mathrm{GTR}}$ model with $\mathrm{CF}_{\mathrm{HKY}}$ and $\hat{\omega}_{\mathrm{CNF}_\times}$ by $\hat{\omega}_{\mathrm{CF}}$.