SI Text

Computing the Fraction of Functional 4GCBs Using Divergence/Polymorphism Comparisons

The most conservative assumption for our purpose is that all functionally significant nucleotides in 4GCBs are associated with the same selection coefficient. If α is the fraction of functional bases among 4GCBs and β is the fraction of neutrally evolving bases among 4GCBs, nucleotide diversity is given by:

$$\pi = \alpha \cdot 4N_e\mu \frac{2N_e s + e^{-2N_e s} - 1}{N_e s\left(1 - e^{-2N_e s}\right)} + \beta \cdot 4N_e\mu \quad \textbf{[1]}$$

and divergence in the human lineage after split from chimpanzee is given by:

$$D = \alpha \cdot 2N_e\mu t \cdot \frac{1 - e^{-2N_e s}}{1 - e^{-4N_e s}} + \beta t\mu \quad \textbf{[2]}$$

Here t is divergence time since the split with chimpanzee, μ is the mutation rate, N_e is the effective population size, and s is the selection coefficient.

Next, the ratio of nucleotide diversity (normalized to diversity in neutrally evolving sites) to divergence (normalized to neutral divergence) is given by:

$$R = \frac{\alpha \cdot \dfrac{2N_e s + e^{-2N_e s} - 1}{N_e s \cdot \left(1 - e^{-2N_e s}\right)} + \beta}{\alpha \cdot 2N_e \dfrac{1 - e^{-2N_e s}}{1 - e^{4N_e s}} + \beta} \quad \textbf{[3]}$$

Given observed R, the minimal estimate of α over all possible values of s is 0.18.

Variation in θ Between Functional and Neutral Conserved Sites

All these estimates are based on the assumption that values of θ ($4N_e\mu$) are identical in functional conserved sites and neutral conserved sites. The following unlikely combination of parameters may theoretically render these estimates nonconservative. Conserved neutral sites should have θ lower than nonconserved sites, whereas conserved functional sites should have value of θ higher than in nonconserved neutral sites. Because CpG dinucleotides were excluded from consideration, variation in mutation rates is unlikely to be >2-fold. Effective population size cannot be consistently higher in functional sites compared with neutral sites (Hill-Robertson interference theoretically can make effective population size lower in functional regions). Therefore, this "parameter conspiracy" cannot change the estimate substantially. Moreover, the estimate would still accurately approximate the fraction of *de novo* deleterious mutations in the genome, a parameter of critical importance for human genetics.

Computing the Fraction of 50-bp Windows Containing Functional Positions

To compare our estimate with genomewide estimates of the fraction of 50-bp windows under selection in the genome, we conservatively assumed that all windows contained either fully functional or fully nonfunctional 4GCBs (that is, functional bases were as tightly clustered as possible). The average number of 4GCBs in a functional versus a nonfunctional 50-bp window is unknown. Based on our observation that sparsely distributed 4GCBs show the same shift in allele frequency as densely clustered 4GCBs, we assume (nonconservatively) that these quantities are the same. Because we estimated ~20% of noncoding 4GCB sites under selection, this meant 20% of 50-bp noncoding windows in the alignable fraction of the genome (27% of the entire genome) were functional; this yields an estimate of 5.5% of all 50-bp noncoding windows genomewide containing functional bases. If, nonconservatively, functional 4GCBs would be

distributed randomly among windows, as many as 26% of all 50-bp noncoding windows genomewide would contain at least one functional 4GCB.