Supporting Text

Uncertainty in the Designation of Ancestral States. Here we allow for uncertainty in the assignment of ancestral and derived states by applying a substitution model. Given the observed nucleotides at a SNP and the nucleotide at the homologous site in an outgroup, our main concern is to compute the probability of observing each of the nucleotides in the most recent common ancestor (MRCA) of the sample. To perform this calculation, we make two main assumptions: (*i*) the coalescence time of the sample (the time back to the MRCA) is much smaller than the divergence time between the sampled species and the outgroup, and (*ii*) for each SNP, only one mutation has occurred since the time of the MRCA; therefore, we assume an infinitely many sites model within species, but a finite sites model between species. Let *a* and *b* be the allelic states for a SNP with two alleles segregating, let *c* be the allelic state at the homologous site in the outgroup, let *m* be the allelic state of the MRCA of the sample, and let t_{div} be the divergence time between the two species. By exponentiating the rate matrix of a substitution model, we can calculate the probability of moving from one state to another over some time $2t_{div}$, *e.g.* $\Pr(c \to a | 2t_{div})$. The probability that *a* is ancestral to *b*, conditional on the existence of an *a/b* polymorphism and state *c* in the outgroup, is

$$\mathbf{Pr}(m=a) = \begin{cases} \dfrac{\theta_{a\to b}\,\mathbf{Pr}(c \to a \mid 2t_{div})}{\theta_{a\to b}\,\mathbf{Pr}(c \to a \mid 2t_{div}) + \theta_{b\to a}\,\mathbf{Pr}(c \to b \mid 2t_{div})} & \text{if } c \text{ is available} \\ \dfrac{\pi_a \theta_{a\to b}}{\pi_a \theta_{a\to b} + \pi_b \theta_{b\to a}} & \text{otherwise} \end{cases} \quad \textbf{[S1]}$$

$\theta_{a\to b}$ and $\theta_{b\to a}$ are the relative mutation rates from *a* to *b* and *b* to *a*, respectively, and π_a and π_b are the stationary probabilities of states *a* and *b*, respectively, which we determine empirically. The above expressions assume that the model is time-reversible.

In order to apply the above expression to the data, we obtained homologous chimp sequences for each gene by BLAT search (1) of the human reference sequence for each gene (available at the NIEHS SNPs web site) against the Broad 1 build of the chimp genome, available at http://genome.ucsc.edu. To calculate $\theta_{a\to b}$, $\theta_{b\to a}$, $\mathbf{Pr}(c \to a \mid 2t_{div})$, and $\mathbf{Pr}(c \to b \mid 2t_{div})$, we used a trinucleotide substitution model (2) that accounts for context-dependent variation in mutation rate common in mammals, such as hypermutability at CpG dinucleotides (2, 3). Hwang and Green (2) estimated the parameters of this model using data from 19 mammalian species. We set $t_{div}=10N$ generations for humans and chimps. For the analysis of indels, we did not use outgroup or mutational bias information.

Computational Details for Evaluating and Optimizing Likelihood Functions. In order to approximate Kimura's expression (Eq. **3**) for the transient distribution of allele frequency, we calculate the first 200 terms of the series. To calculate the Gegenbauer polynomials, we use the `gsl_sf_gegenpoly_n` function of the GNU scientific library (4). Also note that Kimura uses a definition of the Gegenbauer polynomials that are no longer in common usage. Kimura's solution can be written in terms of the modern definition by substituting the relationship given in section 8.932 of ref. 5 into Kimura's (6) equation 4.10, which involves hypergeometric functions. We performed several tests

to determine whether the numerical solution to the diffusion accurately represents the true transient distribution of allele frequency. First, under neutrality, the numerical solution closely matches analytical predictions (expression 4) for all allele frequencies and parameter combinations; the probability density is generally accurate to within 0.1% for all allele frequencies. Numerical and analytical evaluations of likelihood functions (Eq. 10) show a similarly close correspondence. Second, after ~$4N_C$ generations, the numerical solution converges to the stationary distribution of allele frequencies. And finally, Von Neumann stability analysis (7) reveals that the Crank-Nicolson algorithm is unconditionally stable. Likelihood functions were evaluated using the extended midpoint numerical integration algorithm (ref. 7, chapter 4), then maximized using the Fletcher-Reeves-Polak-Ribiere optimization routine (ref. 7, chapter 10).

Estimating Effective Population Sizes and Converting the Time Scale to Years. In order to convert our estimate of the time back to the size change into generations, we (*i*) estimate the unscaled mutation rate using human-chimp divergence data, (*ii*) estimate the scaled (by $4N_C$) mutation rate using human polymorphism data, (*iii*) use estimates from steps *i* and *ii* to estimate the current effective population size, and (*iv*) rescale time into generations. In noncoding regions of the 301 genes we analyzed, 3.86 Mb of homologous chimp sequence was available, and in these regions we observed 44,079 fixed differences (1.14% divergence). Under neutrality, the per-generation substitution rate is equal to the mutation rate and independent of population size (8). Therefore, we estimate the per generation mutation rate of the entire region to be $\hat{\mu}_1 = 0.0735$ (=(44,079 substitutions)/(2*6x10^6 years)/(1 generation/20 years), or 1.90 x 10^{-8} per base pair per generation, assuming a generation time of 20 years and a human-chimp divergence time of 6 million years. Next, using polymorphism data from the regions where we have outgroup information, we estimate the population-scaled mutation parameter θ_1 (=$4N_C\mu_1$) as:

$$\hat{\theta}_1 = \frac{S}{\sum_{m=2}^{n}\left(\frac{S_m}{S}\sum_{j=1}^{m-1}F_1\left(j,m;\hat{\tau},\hat{\upsilon}\right)\right)} = 15,088 \qquad \textbf{[S2]}$$

where S is the total number of SNPs, S_m is the number of SNPs that were sampled in m chromosomes, and n is the maximum sample size. Our estimate of the current effective population size is then $\hat{N}_C = \hat{\theta}_1/\left(4\hat{\mu}_1\right) = 51,340$, and we estimate the ancestral size to be $\hat{N}_A = \hat{\upsilon}\hat{N}_C = 8,211$. Therefore, our estimate of the time back to the size change ($\hat{\tau} = 0.00885$ $2N_C$*generations) translates to 908 generations, or approximately 18,200 years.

1. Kent, W. J. (2002) *Genome Res.* **12,** 656-664.
2. Hwang, D. G. & Green, P. (2004) *Proc. Natl. Acad. Sci. USA* **39,** 13994-14001.
3. Nachman, M. W. & Crowell, S. L. (2000) *Genetics* **156,** 297-304.
4. Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth M., & Rossi, F. (2004) *GNU Scientific Library Reference Manual* (Network Theory Limited, Bristol, U.K.), Version 1.5.

5. Gradshteyn, I. S., & Ryzhik, I. M. (2000) *Table of Integrals, Series, and Products* (Academic, New York), 6th Ed.
6. Kimura, M. (1964) *J. Appl. Prob.* **1,** 177-232.
7. Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vatterling, W. T. (1988) *Numerical Recipes in C* (Cambridge Univ. Press, Cambridge, U.K.), 2nd Ed.
8. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).