Human Genome Ultraconserved Elements Are Ultraselected

Sol Katzman,[1]* Andrew D. Kern,[2]* Gill Bejerano,[2]† Ginger Fewell,[3] Lucinda Fulton,[3] Richard K. Wilson,[3] Sofie R. Salama,[2,4] David Haussler[1,2,4]‡

Unexpectedly long regions of extremely conserved DNA, known as ultraconserved elements, were first found by comparing the human, mouse, and rat genomes (*1*).

Although the DAF spectrum of the nonsynonymous sites is consistent with that observed previously, the spectrum for the ultraconserved sites is qualitatively different (Fig. 1). Large fractions of both the segregating ultraconserved sites (55%) and the nonsynonymous sites (41%) are present in only one allele in one sample. However, only 3% of the segregating ultraconserved sites exhibit DAFs of more than 25%, compared with 14% of the segregating nonsynonymous sites (χ^2 P value of 0.002), even after performing a normalization to a common sample size of 80 chromosomes (*6*).

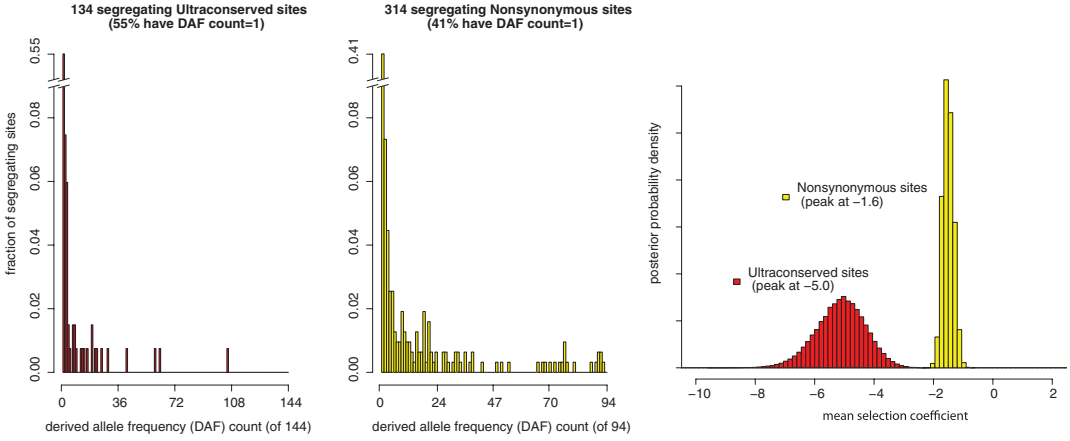

Fig. 1. Ultraconserved elements are under stronger selection than protein-coding regions. (**Left** and **center**) Histograms of the derived allele frequency counts for segregating sites in the indicated categories. In each histogram the first bar, corresponding to singleton heterozygotes (DAF count = 1), is truncated. (**Right**) The Bayesian posterior distributions for the mean selection coefficient. The x axis is given in units of $\alpha = 2N_e s$, where N_e is the effective population size and s is the fitness parameter.

Most are non–protein-coding regions, unique to vertebrates, and have undergone little or no evolutionary change since mammal and bird ancestors diverged about 300 million years ago. Many may function as distal enhancers for neighboring developmental genes (*2*). However, the reason for their extreme conservation remains a mystery. They could be unusually large patches of sites under weak levels of negative selection (*3*, *4*) or simply mutational cold spots.

We measured the derived (new) allele frequency (DAF) spectrum for the segregating human polymorphisms in the ultraconserved regions. It is markedly shifted toward rare derived alleles, as is characteristic of regions under negative selection in which introduced mutations are unlikely to spread to high frequencies within populations.

We analyzed genomic DNA sequences in 72 individuals (a mixture of European Americans and African Americans) spanning 315 of the ultraconserved elements and found 134 segregating sites. We compared the DAFs for these sites with those in 314 segregating nonsynonymous sites in 211 genes obtained from 47 individuals of similar background available from the SeattleSNPs consortium (*5*).

To estimate the distribution of selection coefficients from these DAF spectra, we applied a hierarchical Bayesian model in which the mean selection coefficient for a set of bases is a random variable whose distribution we estimate via Markov chain Monte Carlo (MCMC) methods (*6*). Negative values imply that derived alleles are deleterious. A comparison of the posterior distributions (Fig. 1) shows that the ultraconserved sites are, on average, under purifying selection that is three times greater than that acting on nonsynonymous sites. The 95% credible intervals do not overlap at all.

Such estimates are subject to ascertainment bias, both in the selection of segregating sites (a bias we avoid by completely resequencing the entire region) and implicit in the definition of the ultraconserved regions themselves. A region of the genome containing a segregating site with high DAF is likely to show a difference between the reference human genome and the reference genomes of mouse and rat and hence be excluded from study. Our probability model compensates for such bias (fig. S1), which also applies to polymorphism studies of other conserved regions. In addition, a separate analysis shows that our results are not influenced by different strengths of linkage between sites within the separate classes analyzed (*6*). We can rule out other regional effects because the bases immediately flanking the ultraconserved regions have a much lower mean selection coefficient (fig. S3).

Previous studies have indicated that conserved noncoding regions can exhibit selection coefficients comparable to those of protein-coding regions (*7*). Our analysis shows that selection in the vertebrate-specific ultraconserved noncoding regions is in fact much stronger. These data argue that ultraconserved elements are currently, as well as historically, strongly constrained functional elements.

References and Notes

1. G. Bejerano *et al.*, *Science* **304**, 1321 (2004); published online 6 May 2004 (10.1126/science.1098119).
2. L. A. Pennacchio *et al.*, *Nature* **444**, 499 (2006).
3. P. D. Keightley, M. J. Lercher, A. Eyre-Walker, *PLoS Biol.* **3**, e42 (2005).
4. G. V. Kryukov, S. Schmidt, S. Sunyaev, *Hum. Mol. Genet.* **14**, 2221 (2005).
5. J. M. Akey *et al.*, *PLoS Biol.* **2**, e286 (2004).
6. Materials and methods are available on *Science* Online.
7. J. A. Drake *et al.*, *Nat. Genet.* **38**, 223 (2006).
8. We thank J. Kent, M. Diekhans, D. Thomas, K. Pollard, C. Lowe [University of California Santa Cruz (UCSC)], J. Reed, S. Scott (Genome Sequencing Center), W. Schackwitz, J. Martin, L. Pennacchio (U.S. Department of Energy Joint Genome Institute), P. Robertson (SeattleSNPs), the UCSC Genome Browser group, and anonymous reviewers. Funding was provided by NIH National Human Genome Research Institute (S.K., A.D.K., G.B., D.H., and R.K.W.) and Howard Hughes Medical Institute (S.R.S. and D.H.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/317/5840/915/DC1
Materials and Methods
Figs. S1 to S3
Reference

12 March 2007; accepted 28 June 2007
10.1126/science.1142430

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA. [2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064, USA. [3]Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108, USA. [4]Howard Hughes Medical Institute, University of California, Santa Cruz, CA 95064, USA.

*These authors contributed equally to this work.
†Present address: Department of Developmental Biology and Department of Computer Science, Stanford University, Stanford, CA 94305, USA.
‡To whom correspondence should be addressed. E-mail: haussler@soe.ucsc.edu