ARTICLE

The Strength of Selection on Ultraconserved Elements in the Human Genome

Christina T. L. Chen, Jen C. Wang, and Barak A. Cohen

Ultraconserved elements are stretches of consecutive nucleotides that are perfectly conserved in multiple mammalian genomes. Although these sequences are identical in the reference human, mouse, and rat genomes, we identified numerous polymorphisms within these regions in the human population. To determine whether polymorphisms in ultraconserved elements affect fitness, we genotyped unrelated human DNA samples at loci within these sequences. For all single-nucleotide polymorphisms tested in ultraconserved regions, individuals homozygous for derived alleles (alleles that differ from the rodent reference genomes) were present, viable, and healthy. The distribution of allele frequencies in these samples argues against strong, ongoing selection as the force maintaining the conservation of these sequences. We then used two methods to determine the minimum level of selection required to generate these sequences. Despite the lack of fixed differences in these sequences between humans and rodents, the average level of selection on ultraconserved elements is less than that on essential genes. The strength of selection associated with ultraconserved elements suggests that mutations in these regions may have subtle phenotypic consequences that are not easily detected in the laboratory.

Five percent of the human genome is estimated to be under purifying selection.[1,2] However, only 1.5% of the genome encodes protein, leaving twice as much conserved noncoding DNA as coding DNA. Consistent with this estimate, thousands of conserved noncoding sequences have been discovered in studies that sought to identify mammalian sequences with unusually slow rates of substitution.[3–5] At the extreme end of the sequences identified in these studies are the ultraconserved elements,[6] sequences in which runs of ⩾200 consecutive nucleotides are identical in alignments from the human, mouse, and rat reference genomes.

The underlying assumption of comparative genomics is that sequences that contribute to the fitness of an organism will evolve slowly, relative to selectively neutral sequences. Thus, the ultraconserved elements, which evolve exceptionally slowly, might encode important functions. An alternate hypothesis is that these sequences are situated in regions of the genome with low mutation rates, resulting in fewer than expected nucleotide substitutions over time.

Drake et al.[7] suggested that conserved noncoding sequences are likely to be functional and not mutation cold spots, because the derived alleles in these regions show a bias toward being minor-frequency alleles. On the basis of this observation, Drake et al.[7] concluded that purifying selection maintains these sequences in the genome. Kryukov et al.[8] also concluded that purifying selection, rather than a decrease in mutation rate, drives the conservation of these sequences. Despite the high levels of conservation these sequences exhibit, both Kryukov et al.[8] and Keightley et al.[9] suggested that mutations in conserved noncoding regions are only slightly deleterious. However, the strength of selection required to maintain the sequence conservation of ultraconserved elements, the most extreme representatives of conserved noncoding sequences, has yet to be determined.

In this study, we estimated the magnitude of selection consistent with the maintenance of ultraconserved elements, by analyzing the distribution of polymorphisms within these elements and the nucleotide differences in these sequences in the chimpanzee reference genome. We also compared the estimated selection coefficients with those associated with essential genes, to appreciate their significance. By determining the magnitude of selection that constrains the evolution of ultraconserved elements, we will be better able to devise appropriate experiments that reveal their potential functions.

Material and Methods

Polymorphisms in Ultraconserved Elements

The coordinates of the ultraconserved elements were converted to the May 2004 version (hg17) of the University of California–Santa Cruz (UCSC) Genome Browser.[10] The coordinates of all recorded SNPs in the human SNP database (dbSNP)[11] were checked to see whether they fall within the coordinates of each ultraconserved element. Each SNP that was found in an ultraconserved element was checked to see whether frequency information was recorded, whether it was found using two different methodologies, and whether it was withdrawn after submission.

We calculated the *P* value of observing, at most, 24 validated

From the Department of Genetics, Center for Genome Sciences (C.T.L.C.; B.A.C.), and Department of Psychiatry (J.C.W.), Washington University School of Medicine, St. Louis

Received September 26, 2006; accepted for publication January 25, 2007; electronically published February 20, 2007.

Address for correspondence and reprints: Dr. Barak A. Cohen, Department of Genetics, Center for Genome Sciences, Campus Box 8510, Washington University School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108. E-mail: cohen@genetics.wustl.edu

Am. J. Hum. Genet. 2007;80:692–704. 0002-9297/2007/8004-0011$15.00
DOI: 10.1086/513149

SNPs in these ultraconserved regions, using two approaches. First, we used cumulative Poisson statistics with a genome-average SNP density of 1.84 SNPs per 1 kb of sequence. The genome-average SNP density was derived by dividing the number of all verified SNPs in the database by the size of the human genome. λ in the Poisson equation was computed by multiplying the genome average by the number of bases in ultraconserved elements (126,007 bp). The second approach was to use the empirical frequency distribution of SNPs in the genome. We obtained this distribution by randomly sampling 100,000 different sets of genomic regions that matched the length distribution of the ultraconserved elements and counting the number of validated SNPs in each set.

Selection of SNPs Located Within and Outside Ultraconserved Elements

The two genotyping experiments were approved by the appropriate institutional review boards, and all human DNA samples were deidentified. For the first experiment, we selected 24 SNPs. Half of the SNPs are located within the ultraconserved elements; 9 of 12 lie in intergenic regions, and the remaining 3 are in introns or UTRs. The other 12 SNPs, selected to be controls, were located in regions with a low probability of being under selection. We determined whether a 50-bp window in the human-mouse-rat (HMR) alignment was under selection by calculating the percentage of identity in that window. A "match" occurred when all three species had the same nucleotide in the same position. Otherwise, a "mismatch" was recorded for that position. The percentage of identity for a given window was the total number of match positions divided by the window size. A score was then derived from the percentage of identity by using the scoring system that was modified from previous studies[2] and was based on the cumulative binomial distribution. We modified the scoring scheme by using HMR ancient repeats to estimate the expected frequency of neutral positions with three-way matches, instead of human-mouse ancient repeats. Any window in which >91% of the positions were identical in all three species had a 95% probability of being under selection (C. T. L. Chen and B. A. Cohen, unpublished data). To ensure that the control SNPs we picked were not located in regions of high selection, we scanned 50-bp flanking sequences surrounding each SNP of interest, using 50-bp overlapping windows, and calculated the score of each window. A SNP was selected only if <91% of the bases were three-way matches in all the windows. The distances between each of the paired SNPs range from 28 bp to just over 1 kb. About half of the nearby SNPs had been validated by multiple labs, according to dbSNP, as of July 2005.

We genotyped these SNPs in 752 case-control human DNA samples provided by the Collaborative Study on the Genetics of Alcoholism (COGA) Consortium. One SNP in the first experiment (*rs17049105* in ultraconserved region [UC] 51) had a low frequency of derived alleles, as documented in dbSNP. To ensure that the observation associated with this SNP was not due to inadequate sample size, we genotyped it in additional samples, along with one SNP upstream and one SNP downstream from it. These two SNPs were located outside regions of high conservation. We also chose two SNPs located within one ultraconserved region (UC 268) to genotype in additional samples, since there were no frequency data associated with them in dbSNP. In addition, we selected two more pairs of SNPs to genotype in the second experiment. Each pair included one SNP located within the ultraconserved regions (UC 140 and UC 353) and another SNP located outside regions of high conservation, defined as detailed above. We genotyped these nine SNPs in 721 control human samples provided by the genetic core at Alzheimer Disease Research Center at Washington University. The MassARRAY system was employed in genotyping human DNA samples.[12] SpectroDESIGNER software was used to select primers for each SNP (Sequenom). Standard Sequenom PCR protocols were used, followed by shrimp alkaline phosphatase treatment and the homogenous MassEXTEND reaction, as detailed in the MassARRAY application notes (Sequenom). The SpectroACQUIRE and SpectroAnalyzer modules in the Typer software were used to analyze the SNP data (Sequenom). We compared the allele frequencies of SNPs between the COGA case and control samples and found no significant differences between these two groups of samples. Therefore, we combined the samples in all later investigations.

To determine whether a SNP was in Hardy-Weinberg equilibrium (HWE), we used two different implementations of Fisher's exact test. General goodness-of-fit tests, such as the χ^2 and likelihood (*G* test) tests, were not suitable, since some of the expected genotype numbers were <10, making the sampling distribution of the test statistics only approximately equal to the theoretical χ^2 distribution.[13] The asymptotic assumption did not hold in these cases. The first implementation used the Markov chain–Monte Carlo (MCMC) method to estimate the *P* values and was developed for multiple alleles.[14] The program was run using the parameters: 2,000 initial steps, 500 chunks, and 5,000 as the size of each chunk. The second implementation (EXACT) was derived from the first implementation, with specific application to biallelic SNPs.[15] An α level of .05 was selected as a threshold. Any SNP with $P < .05$ in either test was deemed to be out of HWE. To calculate linkage disequilibrium between pairs of SNPs, the genotype data were compiled to construct haplotype information for each individual. Individuals with ambiguous haplotypes were removed before calculation of linkage disequilibrium. Fisher's exact test was applied to each set of genotype data. An α level of .05 was selected as a threshold. Any pair of SNPs with a *P* value <.05 was considered to be in linkage disequilibrium.

Calculation of Selection Coefficients with the Use of Fixed Differences

To calculate the strength of selection acting on ultraconserved elements, we made two assumptions. Since the probability of observing a long run of consecutive nucleotides was so low,[6] we assumed there were no neutrally evolving positions in ultraconserved elements. Also, since we were interested in the average selection on ultraconserved elements and not in the selection on individual bases, we assumed that each nucleotide was under the same magnitude of selection in these regions. We employed equations developed by Kimura[16,17] to relate the amount of sequence divergence between human and chimpanzee with the strength of the selection coefficient. The substitution rate per nucleotide between two species can be modeled as

$$k = 8N_e\mu\pi(p,s) + 4N_e\mu t\pi(p,s) \ , \qquad (1)$$

where N_e is the effective population size; $\pi(p,s)$ is the fixation probability as a function of s, the selection coefficient, and of p, the initial frequency of the mutant allele; t is the time, in generations, since divergence of the two species; and μ is the mutation rate per base per generation.

Table 1. Definition of Relative Fitness for Possible Genotypes in Kimura's Equations

Genotype[a]	Relative Fitness[b]
A_1A_1	$1 + s$
A_1A_2	$1 + hs$
A_2A_2	1

[a] A_1 refers to the derived allele, and A_2 refers to the ancestral allele.

[b] s is the selection coefficient, and h is the dominance factor.

Nucleotide differences between the human and chimpanzee genomes could have arisen in two ways. Humans and chimpanzees may have inherited different alleles from their common ancestors, as modeled by the first part of equation (1). Mutations could also have occurred after the speciation event and could have subsequently reached fixation, as modeled by the second part of equation (1). The fixation probability of any mutation $\pi(p,s)$ can be calculated in terms of the dominance parameter[17] h:

$$\pi(p,s) = \frac{\int_0^p e^{-2cDx(1-x)-2cx}dx}{\int_0^1 e^{-2cDx(1-x)-2cx}dx}, \qquad (2)$$

where p is the initial mutant-allele frequency, $c = N_e s$, and $D = 2h - 1$.

We solved for s, given h, k, N_e, μ, t, and p, by rearranging equations (1) and (2):

$$\pi(s) = \frac{\int_0^p e^{-2cDx(1-x)-2cx}dx}{\int_0^1 e^{-2cDx(1-x)-2cx}dx} = \frac{k}{4N_e\mu(t + 2N_e)}. \qquad (3)$$

Assuming that the effective population size was the same for both chimpanzee and human, we set N_e to be 10,000.[18] Assuming that any mutation occurring in the ultraconserved region was deleterious, we set μ to be the average mutation rate in the human genome, which was 2.5×10^{-8}.[19] The number of generations since the divergence of human and chimpanzee was estimated to be 250,000, with an average life span of 20 years.[20] Assuming that the mutation leading to a new allele was rare, we set p to be $1/(2N_e) = 1/20{,}000$.

We used mouse as an outgroup, to determine which allele in chimpanzee and human was the ancestral allele. The fitness scheme for possible genotypes—A_1A_1, A_1A_2, and A_2A_2—can be found in table 1, where A_1 refers to the derived allele and A_2 refers to the ancestral allele. Since the real dominance factor, h, was not known, we sampled different values of h that represent different selection models (table 2).

To determine the number of fixed differences in the ultraconserved elements between human and chimpanzee genomes, we compared each of the 481 ultraconserved sequences with the chimpanzee genome (November 2003 version) by using the BLAT tool (UCSC Genome Browser).[2,10] When BLAT did not yield hits, we checked whether the queried sequences were situated in gaps between contigs or supercontigs. We found one ultraconserved element (UC 294) to be completely missing in the chimpanzee assembly. There was one contig (contig 36351) spanning the entire region around UC 294, but BLAT could not identify any sequence on that contig that was homologous to UC 294. We looked for UC 294 in 10 chimpanzee DNA samples, using PCR with primers internal to this element, and showed that this element was present in the chimpanzee even though the sequence was not identified in the November 2003 assembly.

We did not include any bases that were deleted or inserted in the human genome relative to the chimpanzee genome, since many insertions and deletions were likely to be errors in the genome assemblies, such as the UC 294 assembly error described above. The substitution rate was calculated by dividing the number of fixed differences by the total number of aligned bases between the two genomes; it was 1.16×10^{-3}. Assuming that all ultraconserved elements evolved as one allele and that the same selection force was acting on each base in these regions, we solved for $\hat{s}$, using equation (3) with 1.16×10^{-3} as the substitution rate. We then relaxed these assumptions by calculating selection coefficients for each of 481 ultraconserved elements, using the substitution rates derived for each element. We performed the calculations in Mathematica. To interpret the values of selection coefficients, we calculated $\hat{\gamma}$ as $\hat{\gamma} = N_e\hat{s}$ and estimated the ratio of expected numbers of replacement differences between the two species under the selected model and neutral model to be $2\hat{\gamma}/(1 - e^{-2\hat{\gamma}})$ at the present time.[21]

Calculation of Selection Coefficients with the Use of Human Polymorphisms

We employed the Poisson random field framework to model polymorphisms in these sequences and to calculate the maximum-likelihood estimate of selection coefficients.[22] To accommodate different sample sizes for each SNP that we genotyped and to estimate selection coefficients given the dominance factor, several modifications were made.

Let

$$l(\gamma,h\,|\,x) = \ln(n!) - \sum_{i=1}^{n-1}\ln(x_i!) + \sum_{i=1}^{n-1} x_i \ln\left[\frac{F(n,i;\gamma,h)}{\sum_{j=1}^{n-1} F(n,j;\gamma,h)}\right],$$

as defined in the work of Williamson et al.,[22] where $\gamma = 2N_e s$ for diploid organisms, h is the dominance factor, n is the total number of alleles (constant for each SNP), i is the number of alleles with ancestral SNPs (different for each SNP), and x_i is the number of SNP having i copies in n alleles.

Table 2. Definition of Dominance Models

Dominance Factor (h)	Selection Coefficient	
	Positive[a]	Negative[a]
−1	Underdominance	Overdominance
0	A_1 recessive, A_2 dominant	A_1 dominant, A_2 recessive
.5	Incomplete dominance	Incomplete dominance
1	A_1 dominant, A_2 recessive	A_1 recessive, A_2 dominant
2	Overdominance	Underdominance

[a] A_1 refers to the derived allele, and A_2 refers to the ancestral allele.

To optimize $l(\gamma,h|x)$, we needed to optimize

$$\sum_{i=1}^{n-1} x_i \ln\left[\frac{F(n,i;\gamma,h)}{\sum_{j=1}^{n-1} F(n,j;\gamma,h)}\right] .$$

To accommodate different sample sizes for each SNP, we rewrote the above as

$$\sum_{k=1}^{S} \ln\left[\frac{F(n_k,i_k;\gamma,h)}{\sum_{j=1}^{n-1} F(n_k,j_k;\gamma,h)}\right] ,$$

where S is the total number of SNPs, n_k is the total number of alleles for SNP k, and i_k is the number of ancestral alleles that SNP k has.

Instead of optimizing $l(\gamma,h|x)$, we optimized $l(\gamma|h,x)$. This was equivalent to optimizing

$$\sum_{k=1}^{S} \ln\left[\frac{F(n_k,i_k;\gamma,h)}{\sum_{j=1}^{n-1} F(n_k,j_k;\gamma,h)}\right] ,$$

since $l(\gamma,h|x) \propto l(\gamma|h,x)$. The optimization was performed in a C program with the use of different h values, as listed in table 2.

Similar modifications were done to the equations to calculate the 95% CI for $\hat{\gamma}$. We first calculated $\hat{\gamma}$ for the ultraconserved elements as a single allele and then calculated $\hat{\gamma}$ only for those elements that harbored SNPs.

In this analysis, mouse was not an appropriate outgroup to use to determine which allele in the human population was the ancestral allele. Mutations in these elements could have occurred and fixed on the lineage leading to humans, after humans and rodents diverged. Instead, we used the chimpanzee as an outgroup. The fitness scheme can be found in table 3.

Calculations of Probabilities of Observing at Least 12 Frequent SNPs under Weak and Strong Selection

We compared the probabilities of observing at least 12 frequently derived SNPs under weak and strong selection, given a range of the total number of SNPs within the ultraconserved elements in the population, using all possible allele-frequency distributions of SNPs and a range of dominance models.

First, we estimated the probability of a SNP found at each frequency in a population of 1,000, using

$$\frac{F(n,i;\gamma,h)}{\sum_{j=1}^{n-1} F(n,j;\gamma,h)} ,$$

as defined in the work of Williamson et al.,[22] where n was the sample size, i was the frequency of the derived SNP in the sample, γ was the selection coefficient, and h was the dominance factor. A population size of 1,000 was appropriate, since all SNPs in our sample were genotyped in >1,000 individuals. Since very weak selection has been defined as $|\gamma| \leq 1$, and strong selection has been defined as $|\gamma| \gg 1$,[23,24] we chose $\gamma = -1$ and -5 to represent weak and strong purifying selection, respectively, in our analysis. As before, we used a range of dominance models: $h = -1$, 0, 0.5, 1, and 2.

We defined a SNP as "frequent" if its frequency in the population is ⩾5%. The combinatorial blowup of considering all possible arrangements of SNPs with frequencies from 1% to 99% in ⩾12 different positions necessitated this simplification. Then, the probability of a SNP being rare was computed as the sum of probabilities of a SNP being found in <5% of the individuals. Since the total number of SNPs in the ultraconserved regions was not known, we used 24–100 total SNPs in our calculations. The probability of observing at least 12 frequent SNPs was calculated using the cumulative binomial distribution.

Table 3. Definition of Relative Fitness for Possible Genotypes Employed in the Poisson Random Field Model

Genotype[a]	Relative Fitness[b]
A_1A_1	$1 + 2s$
A_1A_2	$1 + 2hs$
A_2A_2	1

[a] A_1 refers to the derived allele, and A_2 refers to the ancestral allele.

[b] s is the selection coefficient, and h is the dominance factor.

Comparison of Essential Genes with the Ultraconserved Elements

We used the mammalian phenotype browser at Jackson Laboratory to select exons that, when replaced with null alleles, lead to embryonic lethality during fetal growth or development. The human homologues of these exonic sequences were identified, and the nucleotide sequences were used as queries to identify the homologous sequences in the chimpanzee genome with the use of the BLAT tool (UCSC Genome Browser).[10,25] We removed any base that was aligned to a gap in either of the two genomes before counting the number of different bases between the genomes. Since we assumed that there were no neutral bases in the ultraconserved elements, we removed the third position of every codon in the essential genes, to make certain that each base in the essential genes was under selection. For each ultraconserved element and essential gene that was not polymorphic, we calculated the selection coefficient, using equation (3), as detailed earlier with $h = 0.5$, and plotted the distributions of selection coefficients in diffusion time scale ($\hat{\gamma} = N_e s$). We used the Mann-Whitney test to compare the distributions of selection coefficients between ultraconserved elements and essential genes. Lists of essential genes and ultraconserved elements can be found in tables A3 and A4. To assess the overrepresentation of nonexonic ultraconserved elements with high frequencies of derived alleles, we used the hypergeometric distribution.

Results

We investigated whether changes in the nucleotide sequences of ultraconserved elements are tolerated by examining the distribution of verified SNPs in these regions in the human genome. At the time when ultraconserved elements were found, only six validated SNPs were re-

Table 4. Frequencies of SNPs in the Ultraconserved Regions

SNP	Ultraconserved Element	No. of People: In Total Sample	No. of People: With Homozygous Ancestral Alleles	No. of People: With Heterozygous Alleles	No. of People: With Homozygous Derived Alleles[a]	*P* for HWE: MCMC[b] (SD)	*P* for HWE: EXACT[c]
rs17049105	51	1,361	1,145	215	1 (.080)	.0015 (6.45×10^{-5})	.0015
rs1861100	53	728	41	288	399 (.746)	.28 (1.79×10^{-3})	.28
rs10496382	67	726	653	69	4 (.053)	.14 (6.14×10^{-4})	.14
rs13020355	82	709	502	186	21 (.161)	.49 (1.25×10^{-3})	.49
rs2056116	140	701	249	341	111 (.402)	.75 (1.36×10^{-3})	.81
rs2056117	140	681	294	310	77 (.341)	.73 (1.32×10^{-3})	.80
rs17291131	211	714	521	180	13 (.144)	.54 (1.08×10^{-3})	.65
rs12981	268	723	533	178	12 (.140)	.54 (1.09×10^{-3})	.64
rs3902936	268	682	NP	NP	NP	NP	NP
rs3902937	268	687	NP	NP	NP	NP	NP
rs7092999	295	729	237	376	116 (.417)	.11 (1.59×10^{-3})	.11
rs2111796	353	686	NP	NP	NP	NP	NP
rs9572903	353	725	499	210	16 (.167)	.35 (1.33×10^{-3})	.35
rs7143938	374	729	323	338	68 (.325)	.13 (1.60×10^{-3})	.15
rs4300725	433	726	355	312	59 (.296)	.37 (2.05×10^{-3})	.42
rs11573440	461	620	NP	NP	NP	NP	NP

NOTE.—NP = not polymorphic.

[a] Values in parentheses are derived-allele frequencies in the sample.

[b] Uses the MCMC method to estimate the *P* value for the null hypothesis that two alleles are in HWE.

[c] Uses implementations to estimate the *P* values specifically for the biallelic SNPs.

ported in these sequences.[6] In our study, we found 102 SNPs recorded in dbSNP,[11] 24 of which were verified by two or more research groups. Two approaches were used to determine the significance of observing, at most, 24 SNPs in the ultraconserved elements. With a background density of 1.84 validated SNPs per 1,000 nucleotides in the human genome, we calculated the probability of observing, at most, 24 SNPs in the ultraconserved elements to be 1.14×10^{-68}, using cumulative Poisson statistics. With the assumption that all 102 SNPs are validated, the *P* value increases to 2.74×10^{-22}, which remains significantly lower than the genome average. We also generated the empirical frequency distribution of validated SNPs in the genome by randomly sampling 100,000 sets of genomic regions that matched the length distribution of ultraconserved elements and counting the number of validated SNPs in each set. The number of SNPs in each set ranged from 140 to 341, with an average of 203. This suggests that the probability of observing, at most, 24 SNPs in the ultraconserved elements is $<10^{-5}$. With the assumption that all 102 SNPs are validated, the *P* value remains $<10^{-5}$. The paucity of SNPs in ultraconserved regions is consistent with the high conservation of these sequences between humans and rodents.

Ancestral Alleles of Ultraconserved Elements Are Not Required for Normal Development in Humans

If the perfect conservation of ultraconserved elements across species is due to purifying selection, and if mutations in these sequences are deleterious, then some of the polymorphisms in these sequences in human populations may be deleterious recessive mutations. This hypothesis predicts that, given the frequency of a SNP in an ultraconserved region, there should be an excess of heterozygotes and a corresponding shortage of derived-allele homozygotes. Ancestral alleles of ultraconserved elements are defined as alleles that are found in the rodent reference genomes, whereas the derived alleles are alleles that are different from those in the rodents.

To determine whether strong purifying selection acts on the derived alleles, we genotyped a random sample of >600 phenotypically normal, unrelated humans from two sets of SNPs: one composed of 16 SNPs in ultraconserved elements and another set of 16 SNPs located in the neutral regions flanking the ultraconserved elements (table A1). The distributions of heterozygotes between these distinct sets of SNPs were compared. Four of the 16 SNPs in the ultraconserved elements were not polymorphic in the sampled population (table 4). For each of the 12 polymorphic SNPs that lie in an ultraconserved element, we found at least one individual homozygous for the derived allele. Derived-allele homozygotes therefore do not cause embryonic lethality and do not necessarily show gross observable phenotypic abnormalities.

We hypothesized that, if ultraconserved elements are currently under strong selection, we might be able to detect it by testing whether the ancestral and derived alleles are in HWE. If homozygous derived alleles cause embryonic lethality or confer survival disadvantages compared with homozygous ancestral alleles, then the frequencies of the alleles in the sampled population would deviate from HWE. Using Fisher's exact test, we observed only one SNP (*rs17049105*) out of HWE (table 4). For this SNP, the number of individuals with the derived alleles was fewer than expected, which implies that purifying selection may be currently acting on this locus.

To examine whether the SNPs in ultraconserved elements are more likely to be out of HWE, we examined the genotype distributions of SNPs adjacent to the ultraconserved regions. Only 12 of the 16 SNPs genotyped were included in the final analysis. We identified individuals with two copies of the derived allele for all 12 SNPs except one (*rs17195476*) (table A2). However, this SNP, *rs17195476*, was determined to be in HWE, suggesting that the observed lack of individuals with homozygous derived alleles is due to the low frequency of the derived allele. Of 12 SNPs tested, only 1 (*rs471578*) was not in HWE. This is approximately the same rate of occurrence as the SNPs within the ultraconserved elements. Thus, SNPs in ultraconserved elements are not more likely to be out of HWE than are SNPs outside these regions. Our data argue against strong, ongoing selection on ultraconserved regions but do not rule out the possibility that weak selection acts on these elements and maintains their high conservation.

Fixed Differences Are Found in Ultraconserved Regions between the Human and Chimpanzee Genomes

We next investigated whether ultraconserved elements tolerate changes by examining their homologues in the chimpanzee genome. With the assumption that the ultraconserved elements are maintained by purifying selection, any functions of these elements will likely be the same in chimpanzees and humans. We therefore expected that these sequences would be perfectly conserved in the chimpanzee genome, as they are in the human, mouse, and rat genomes. Using BLAT[2] to identify homologues of all 481 ultraconserved elements in the reference chimpanzee genome, we found there were 141 base differences among the 121,830 aligned bases. The average substitution rate was 1.16×10^{-3} substitutions per base, ~10-fold lower than the average rate of substitution between human and chimpanzee. Even this low rate of substitution was unexpected, given that ultraconserved elements were identified because they lacked any substitutions among the reference human, mouse, and rat genomes.

Ultraconserved Elements Are under Negative Selection

We sought to determine what level of selection is consistent with the observations that ultraconserved elements exhibit both polymorphisms within the human population and fixed differences between humans and chimpanzees. Because our objective was to calculate the average magnitude of selection on ultraconserved elements and not on individual nucleotides, we treated each nucleotide in the elements as being under the same level of selection. We calculated the strength of selection, using two methods. First, we estimated selection coefficients, using the number of fixed differences between human and chimpanzee genomes across the entire set of ultraconserved elements. Using mouse as an outgroup, we defined the ancestral allele as the one identical to the allele in the mouse reference genome. We employed Kimura's equations,[16,17] making the assumption that any mutation in these elements that occurred after speciation of the human and chimpanzee had sufficient time to either become fixed or disappear. Since the magnitude of the selection coefficient was confounded with the mode of interactions between two different alleles, we chose five dominance models representing different modes of interactions and calculated the selection coefficient for each model (table 5). Under all dominance models, the selection coefficients (γ) are negative and range from -2.72 to -1.11, which agrees with the hypothesis that the derived alleles of the ultraconserved elements are deleterious. These estimates of selection coefficients represent the minimum amount of selection required on each site in every generation to maintain the observed sequence conservation. To appreciate the strength of selection on the ultraconserved elements, we compared the number of mutations that are expected to become fixed under each estimated selection coefficient with that under the neutral model ($\gamma = 0$). If the derived allele is recessive and the ancestral allele is dominant ($h = 0$), then the number of mutations that would become fixed is 20-fold lower than if the sequences are evolving under the neutral model (table 5). Alternatively, if the ancestral allele is recessive and the derived allele is dominant ($h = 1$), then this ratio becomes sevenfold.

We also estimated selection coefficients by using the Poisson random field framework[22] that incorporates the frequencies of the SNPs in the ultraconserved elements. Mouse was not an appropriate outgroup to use to determine which allele in the human population was the ancestral allele, since mutations in these elements could have occurred and fixed on the lineage leading to humans after the split with rodents. Instead, we used chimpanzee as the outgroup for this analysis. However, in all positions that are polymorphic in humans, the chimpanzee alleles were the same as the rodent alleles. With use of this framework, γ ranged from -3.53 to 0.74 (table 5). The variances of these estimates are large, because of the small number of available SNPs in these regions. Thus, two independent calculations both suggest that, under most dominance models, the derived alleles are slightly deleterious.

Our calculations, based on the observed number of frequent SNPs in ultraconserved elements, suggest that these sequences are under weak selection. Because our estimates are based on genotyping known SNPs and not on exhaustive resequencing of ultraconserved elements, it is possible that an unknown number of SNPs have been missed in our sample. The number and frequency distribution of these missed SNPs could affect our estimates of the implied selection coefficients in these regions. We therefore computed the probability of observing at least 12 frequent SNPs (the number of frequent SNPs we observed in our genotyping experiments) under a model of either weak or strong selection, assuming that there may be >12 actual SNPs in our sample. The result of these cal-

culations suggests that weak selection, rather than strong selection, is the correct model over a very broad range of total possible SNPs (fig. 1).

For example, we observed 12 frequent SNPs in our genotyping experiments. dbSNP contains an additional seven validated SNPs at known frequencies and five more validated SNPs at unknown frequencies in these sequences. The total number of SNPs is, therefore, likely to be at least 24. In this range, strong selection ($\gamma = -5$) is incompatible with our observations, and the likelihood ratio between the two models suggests that weak selection ($\gamma = -1$) is at least 20 times more likely to be the correct model than strong selection, with the assumption of an intermediate dominance model. Overall, we take this analysis as evidence that weak selection, rather than strong, is operating on ultraconserved elements. The analysis does not rule out strong selection completely, especially if the total number of unobserved SNPs is very large. We think that is unlikely to be the case because the observed number of SNPs in ultraconserved sequences is sixfold lower than the average across the genome. Weak selection is also consistent with our estimates from the analysis of substitutions in these regions between humans and chimpanzees.

Genes whose products are essential for the proper development of an organism are assumed to be under strong purifying selection. We compared the strength of selection acting on the ultraconserved elements with that acting on essential genes (tables A3 and A4). If the strength of selection acting on ultraconserved elements is similar to that acting on essential genes, then this would support the notion that the exceptionally high conservation of these sequences reflects important, but currently unknown, functions in the mammalian lineages. We calculated the strength of selection acting on each essential gene and on each individual ultraconserved element, using an additive model ($h = 0.5$), and compared the two distributions of selection coefficients (fig. 2). As expected, purifying selection appears to be acting on all essential genes. The distributions of selection coefficients for ultraconserved elements and essential genes are significantly different ($P < .0001$; Mann-Whitney's rank sum test). The results suggest that, on average, the magnitude of puri-

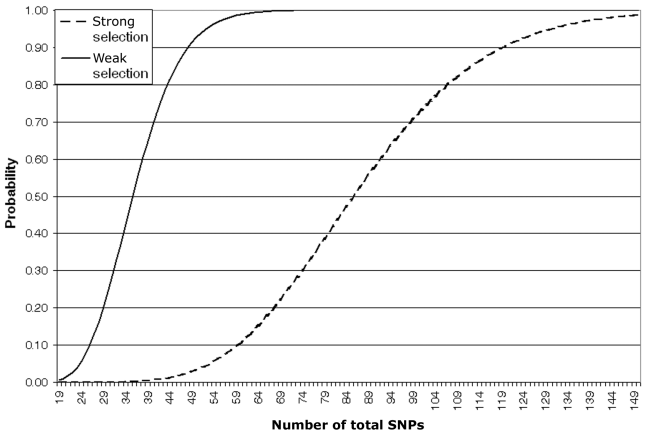

Figure 1. Comparison of weak and strong selection, with the assumption of an intermediate dominance model ($h = 0.5$). Weak and strong selections were defined as $\gamma = -1$ and $\gamma = -5$, respectively. Each point represents the probability of observing at least 12 frequent derived-allele SNPs, with the assumption of different numbers of total (observed and unobserved) SNPs.

Table 5. Strength of Selection Coefficients on the Ultraconserved Elements Calculated for Different Dominance Models

	Kimura Model[b]		Poisson Random Field Model[d]		
Dominance Factor (h)[a]	Selection Coefficient ($\hat{\gamma}$)	(Fixation under $\gamma = 0$)/(Fixation under $\hat{\gamma}$)[c]	Selection Coefficient ($\hat{\gamma}$)	95% Confidence Limits for $\hat{\gamma}$[d]	(Fixation under $\gamma = 0$)/(Fixation under $\hat{\gamma}$)[c]
−1	−2.72	42.19	−3.53	−4.62, −2.44	164.74
0	−2.24	19.46	−.8	−2.10, .48	2.47
.5	−1.91	11.68	1.25	−3.21, 5.70	.37
1	−1.58	7.14	1.13	−1.51, 4.14	.40
2	−1.11	3.70	.74	−1.26, 2.73	.52

[a] Let A_1 be the derived allele, A_2 be the ancestral allele, and $w(x)$ be the fitness of individuals with genotype x. When $\hat{\gamma} < 0$, $h = -1$ implies that $w(A_1A_2)$ is the highest among the three genotypes, $h = 0$ implies that $w(A_1A_1) < w(A_1A_2) = w(A_2A_2)$, $h = 0.5$ implies that $w(A_1A_1) < w(A_1A_2) < w(A_2A_2)$, $h = 1$ implies that $w(A_1A_1) = w(A_1A_2) > w(A_2A_2)$, and $h = 2$ implies that $w(A_1A_2)$ is the lowest among the three genotypes. When $\hat{\gamma} > 0$, $h = -1$ implies that $w(A_1A_2)$ is the lowest among the three genotypes, $h = 0$ implies that $w(A_1A_1) > w(A_1A_2) = w(A_2A_2)$, $h = 0.5$ implies that $w(A_1A_1) > w(A_1A_2) > w(A_2A_2)$, $h = 1$ implies that $w(A_1A_1) = w(A_1A_2) > w(A_2A_2)$, and $h = 2$ implies that $w(A_1A_2)$ is the highest among the three genotypes.

[b] Calculated using fixed differences between chimpanzee and human ultraconserved elements.

[c]

$$\frac{\text{Fixation under } \gamma = 0}{\text{Fixation under } \hat{\gamma}} = \frac{1 - e^{-2\hat{\gamma}}}{2\hat{\gamma}},$$

with the assumption that the mutation rates are the same under both the neutral and the selected model.

[d] Confidence limits calculated using 19 validated SNPs in human ultraconserved elements.

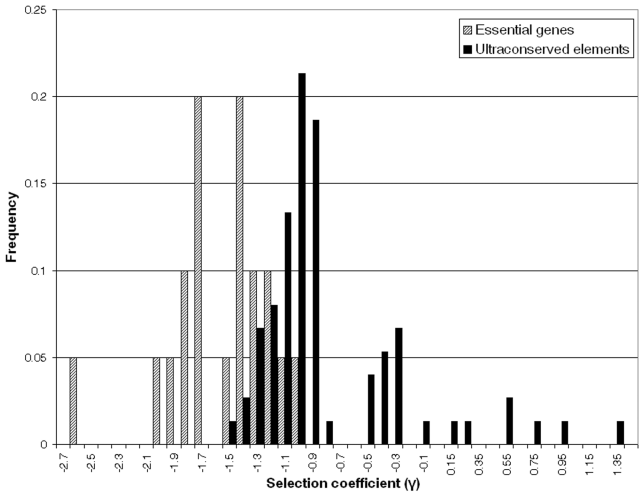

Figure 2. Distributions of selection coefficients for ultraconserved elements and essential genes.

fying selection acting on the ultraconserved elements is weaker than that acting on the essential genes.

We also observed that most of the fixed differences between humans and chimpanzees are concentrated in a small set of ultraconserved elements. Their selection coefficients (γ) range from -1.45 to 1.35, with a median of -0.95. We were unable to distinguish statistically whether this small group of elements represents a distinct set of sequences evolving under different selective constraints or whether they are simply the tail end of a single distribution of selection coefficients encompassing all ultraconserved sequences (data not shown).

For ultraconserved elements that are polymorphic in the human population but show no fixed differences with their chimpanzee counterparts, we calculated selection coefficients, using the Poisson random field model (table 6). Of the 13 elements, 10 contain SNPs with derived alleles at high frequencies. Of those 10 elements, 8 do not overlap exons in humans.[6] We examined the remaining two elements and found no evidence of overlapping known exons in humans. This result suggests an overrepresentation of nonexonic ultraconserved elements with derived alleles at high frequencies ($P < .002$; hypergeometric distribution).

Discussion

The absence of absolute sequence conservation between ultraconserved elements and their homologues in chimpanzees is unexpected. Since the evolutionary distance between chimpanzees and humans is much shorter than that between rodents and humans, one might expect that the sequences of these elements would be preserved in the chimpanzee genome. However, not all ultraconserved elements are conserved in the reference chimpanzee genome. If purifying selection preserves these sequences through evolution, then the functions of these elements may differ significantly in the chimpanzee. Alternatively, the existence of fixed nucleotide differences between chimpanzees and humans in these ultraconserved elements could be a product of past population fluctuations. Studies have suggested that both chimpanzee and human populations experienced a bottleneck in which the population sizes decreased significantly and then rapidly expanded.[26,27] The decrease in population size in both species would allow several slightly deleterious mutations to become fixed, producing a high number of fixed differences between the two species.[8] The subsequent rapid expansion in population size would likely result in few polymorphic alleles within each species.

The most striking feature of the ultraconserved elements is the presence of so many consecutive conserved nucleotides. There are no known functional sequence elements that require such long stretches of specific sequence. Nucleotide alignments of ORFs, noncoding RNAs, and transcription-factor binding sites all show characteristic patterns of substitutions.[28–30] The ultraconserved elements may therefore represent a new class of slowly evolving

Table 6. Selection Coefficients for Ultraconserved Elements That Are Polymorphic in Human Populations but Show No Substitutions with Chimpanzee Homologues

Ultraconserved Element	Type[a]	Selection Coefficient ($\hat{\gamma}$ at $h = .5$)[b]	(Fixation under $\gamma = 0$)/(Fixation under $\hat{\gamma}$)[c]
140	n	35.2	70.4
295	n	2.60	5.22
374[d]	p	1.93	3.95
302	n	1.86	3.81
433	n	1.72	3.55
334	n	.984	2.29
353	n	.538	1.63
82	n	.458	1.53
211	n	.242	1.26
268[d]	p	.204	1.22
440	n	−.459	.610
368	n	−3.20	.0107
269[e]	n	<−5	.000454

[a] As defined by Bejerano et al.[6] Type "e" suggests the element overlaps the mRNA of a known human protein-coding gene (including the UTR regions). Type "n" indicates there is no evidence of transcription of this element from any matching EST or mRNA from any species. Type "p" refers to elements where evidence of transcription is inconclusive.

[b] Scaled s by effective population size. Let A_1 be the derived allele, A_2 be the ancestral allele, and $w(x)$ be the fitness of individuals with genotype x. When $\hat{\gamma} < 0$, $h = 0.5$ implies that $w(A_1A_1) < w(A_1A_2) < w(A_2A_2)$. When $\hat{\gamma} > 0$, $h = 0.5$ implies that $w(A_1A_1) > w(A_1A_2) > w(A_2A_2)$.

[c]
$$\frac{\text{Fixation under } \gamma = 0}{\text{Fixation under } \hat{\gamma}} = \frac{1 - e^{-2\hat{\gamma}}}{2\hat{\gamma}},$$
with the assumption that the mutation rates are the same under both the neutral and the selected model.

[d] Although these elements are classified as "partially exonic" by Bejerano et al.,[6] we did not find any evidence of their overlapping known exons in humans.

[e] Program does not converge.

sequences that are constrained at every position. However, since we calculated average selection coefficients, our data do not rule out the possibility that different positions in these elements are under different levels of selection and that some positions may be neutral with respect to fitness. If ultraconserved elements do contain some neutral positions that are conserved solely by chance, then it is likely that these elements will fall into known functional sequence classes. Indeed, one ultraconserved element was recently shown to be a transcriptional enhancer.[31] The estimation of the number of neutral bases within ultraconserved elements is an ongoing effort that involves comparing the patterns of substitution in multiple mammalian lineages.

Our methods for estimating the average selection acting on ultraconserved elements may be affected by biases within dbSNP. There may be more SNPs located in these elements that have yet to be found and documented. In addition, our approach of counting only fixed nucleotide differences between human and chimpanzee reference genomes does not take into consideration the possible existence of insertions or deletions in ultraconserved regions. It is also possible that some of these fixed differences may, in fact, be polymorphisms in the chimpanzee population. Taken together, these caveats suggest that our estimates of selection may be based on underestimates of polymorphism in these regions. We therefore argue that our estimates are conservative upper bounds of the strength of selection on these elements, since incorporating additional sequence changes into our calculations would likely lower our estimates even further. Moreover, some fraction of the rare polymorphism we observed could result from a past population bottleneck and subsequent expansion. If this scenario were true, then the actual levels of selection on ultraconserved elements are likely to be even weaker than we estimated because some of the polymorphism we observed is a result of rapid population expansion rather than of purifying selection.

The estimates of selection we have calculated refer to the average level of selection across each ultraconserved element. The estimation of selection coefficients for individual bases within ultraconserved elements would take alignments drawn from hundreds of mammalian genomes.[32] To circumvent this problem, population genetic studies often assume an a priori distribution of selection coefficients, usually drawn from the gamma distribution. Since the functions of these elements are unknown, we did not know whether gamma would be an appropriate a priori distribution. We have therefore limited our conclusions to the average levels of selection on ultraconserved elements.

Our results show that selection coefficients as low as 0.0272% per generation can drive nucleotide differences to fixation on the lineages leading to humans and chimpanzees and can maintain the sequence conservation of ultraconserved elements in humans. This magnitude of selection is in agreement with a recent finding that very weak selection appears to be acting on the conserved noncoding regions, defined by the comparison of human and mouse reference genomes.[8] This relatively low level of selection is not enough to drive the allele frequencies of polymorphisms out of HWE in surviving adult populations. Although this estimate of selection coefficients is dependent on the mode of interaction between the ancestral and derived alleles, the detection of phenotypes of individuals with derived alleles may be difficult in a laboratory setting. The feasibility of detecting phenotypic changes relies on adequate sample size and the number of generations in the study. For most studies in the laboratory, a selection of 0.1% is already below the current level of detection.[33] In the case where the minimum selection we calculated in this study acts constantly in every generation, we should not necessarily expect to observe obvious phenotypic changes in individuals with mutations in ultraconserved elements.

Acknowledgments

We thank Alison Goate, the COGA Consortium (funded by National Institute on Alcohol Abuse and Alcoholism and National Institute on Drug Abuse grant U10AA0840), the Alzheimer Disease Research Center at Washington University (funded by National Institutes of Health grant P50AG05681 to Dr. John Morris), and Anne Bowcock, for providing human and chimpanzee DNA samples. We also thank Quo-Shin Chi, for advice on solving Kimura's equations; Scott Williamson, for access to his computer programs and suggestions on how to modify them; Stan Sawyer, for advice on statistical analysis; Rob Mitra, Gil Bejerano, and members of the Cohen Lab, for helpful discussions; and Ed Esparza, for proofreading the manuscript.

Appendix A

Table A1. Linkage Relationship between Each Pair of SNPs inside and outside the Ultraconserved Region

Ultraconserved Region and SNP Inside	SNP outside Ultraconserved Region	Location of Outside SNP Relative to Inside SNP	Two-Tailed P[a]
51:			
rs17049105	*rs13002041*	69 bp downstream	3.98×10^{-11}
rs17049105	*rs2125925*	1,268 bp upstream	1.78×10^{-11}
rs17049105	*rs3886275*	733 bp downstream	3.27×10^{-11}
53:			
rs1861100	*rs12464082*	567 bp upstream	NA[b]
67:			
rs10496382	*rs17029579*	80 bp upstream	1.68×10^{-2}
82:			
rs13020355	*rs12622351*	262 bp downstream	2.50×10^{-6}
140:			
rs2056116	*rs2670760*	390 bp upstream	1.53×10^{-5}
rs2056117	*rs2621243*	602 bp upstream	5.34×10^{-10}
211:			
rs17291131	*rs17310268*	663 bp downstream	1.53×10^{-129}
268:			
rs12981	*rs10985794*[c]	211 bp upstream	1.63×10^{-1}
295:			
rs7092999	*rs2489029*	54 bp downstream	1.28×10^{-150}
353:			
rs9572903	*rs17195476*	152 bp upstream	5.97×10^{-5}
rs2111796	*rs471578*	797 bp downstream	NA[b]
374:			
rs7143938	*rs7150986*[c]	1,159 bp downstream	1.18×10^{-1}
433:			
rs4300725	*rs1365463*	28 bp downstream	4.79×10^{-208}
461:			
rs11573440	*rs11573439*	166 bp upstream	NA[b]

[a] Null hypothesis: two SNPs are at linkage equilibrium.

[b] NA = not available. One or both SNPs in the pair are not polymorphic, and the linkage relationship cannot be determined.

[c] These pairs of SNPs are not in linkage disequilibrium, and the SNPs outside the conserved regions in these pairs are excluded from the analysis.

Table A2. Frequency of SNPs outside the Ultraconserved Elements

		No. of People				P for HWE	
SNP	Nearby Ultraconserved Element	In Total Sample	With Homozygous Ancestral Alleles	With Heterozygous Alleles	With Homozygous Derived Alleles[a]	MCMC[b] (SD)	EXACT[c]
rs3886275	51	681	292	302	87 (.349)	.50 (2.01×10^{-3})	.56
rs2125925	51	667	85	293	289 (.653)	.44 (2.07×10^{-3})	.44
rs13002041	51	728	101	339	288 (.628)	.87 (7.52×10^{-4})	.94
rs12464082	53	733	NP	NP	NP	NP	NP
rs17029579	67	698	8	120	570 (.903)	.39 (8.96×10^{-4})	.52
rs12622351	82	717	491	204	22 (.173)	.90 (3.70×10^{-4})	.90
rs2670760	140	727	600	119	8 (.0928)	.38 (9.68×10^{-4})	.38
rs2621243	140	682	592	88	2 (.0674)	.76 (3.71×10^{-4})	.76
rs17310268	211	688	552	129	7 (.104)	1 (0)	1
rs10985794[d]	268	695	666	26	3 (.0230)	.00089 (7.55×10^{-5})	.0043
rs2489029	295	704	135	372	197 (.544)	.079 (1.38×10^{-3})	.094
rs471578	353	647	462	157	28 (.165)	.0023 (1.04×10^{-3})	.0040
rs17195476	353	733	658	75	0 (.0511)	.25 (5.29×10^{-4})	.25
rs7150986[d]	374	734	717	16	1 (.0123)	.1 (3.22×10^{-4})	.10
rs1365463	433	661	6	116	539 (.904)	.82 (3.89×10^{-4})	1
rs11573439	461	701	NP	NP	NP	NP	NP

NOTE.—NP = not polymorphic.

[a] Values in parentheses are derived-allele frequencies in the sample.

[b] Uses the MCMC method to estimate the P value for the null hypothesis that two alleles are in HWE.

[c] Uses implementations to estimate the P values specifically for the biallelic SNPs.

[d] These SNPs are excluded from analysis since they are not in linkage disequilibrium with the SNPs located within the ultraconserved regions.

Table A3. Essential Genes and Their Selection Coefficients

Gene	GenBank Accession Number	Gene Description	Location[a]	Selection Coefficient ($\hat{\gamma}$ at $h = .5$)[b]
APAF1	NM_013229	Apoptotic peptidase activating factor 1	Chr12: 97541545–97631672	−1.30
C1GALT1	NM_020156	Core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase, 1	Chr7: 7047128–7057219	−1.02
CITIED2	NM_006079	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	Chr6: 139735091–139737478	−1.10
EDG1	NM_001400	Endothelial differentiation, sphingolipid G-protein-coupled receptor	Chr1: 101414596–101419094	−1.83
EPO	NM_000799	Erythropoietin	Chr7: 99963073–99965972	−1.36
IKBKB	NM_001556	Inhibitor of kappa light polypeptide gene	Chr8: 42247985–42309122	−1.72
LIG4	NM_002312	DNA ligase IV	Chr13: 107657793–107665883	−1.32
MEN1	NM_130804	Menin isoform 1	Chr11: 64327571–64335342	−1.68
MGAT2	NM_001015883	Mannosyl (alpha-1,6-)-glycoprotein	Chr14: 49157238–49159948	−1.17
MTF1	NM_005955	Metal-regulatory transcription factor 1	Chr1: 37948943–37994324	−1.82
MYB	NM_005375	V-myb myeloblastosis viral oncogene homolog	Chr6: 135544145–135582002	−1.43
NCOR1	NM_006311	Nuclear receptor corepressor 1	Chr17: 15875983–16059570	−1.97
NDST1	NM_001543	N-deacetylase/N-sulfotransferase 1	Chr5: 149880622–149917965	−1.92
NF1	NM_000267	Neurofibromin 1	Chr17: 26446242–26725590	−2.62
NFAT5	NM_173214	Nuclear factor of activated T-cells 5	Chr16: 68156497–68296054	−1.15
PTHR1	NM_000316	Parathyroid hormone receptor 1	Chr3: 46894239–46920290	−1.66
RCE1	NM_005133	Prenyl protein peptidase	Chr11: 66367458–66370579	−1.73
SERPINC1	NM_000488	Serpin peptidase inhibitor, clade C, member 1	Chr1: 170604598–170618130	−1.49
TULP3	NM_003324	Tubby like protein 3	Chr12: 2870293–2920560	−1.43
VEGFC	NM_005429	Vascular endothelial growth factor C	Chr4: 177979839–178089044	−1.41

NOTE.—Let A_1 be the derived allele, A_2 be the ancestral allele, and $w(x)$ be the fitness of individuals with genotype x. When $\hat{\gamma} < 0$, $h = 0.5$ implies that $w(A_1A_1) < w(A_1A_2) < w(A_2A_2)$. When $\hat{\gamma} > 0$, $h = 0.5$ implies that $w(A_1A_1) > w(A_1A_2) > w(A_2A_2)$.

[a] These coordinates refer to the gene positions in the Human May 2004 (hg17) assembly on the UCSC Genome Browser. Chr = chromosome.

[b] Scaled s by effective population size.

Table A4. Ultraconserved Elements That Show Substitutions with Chimpanzee Homologues

Ultraconserved Element	Type[a]	Selection Coefficient ($\hat{\gamma}$ at $h = .5$)[b]
7	p	−.463
15	n	−.988
16	p	−.335
21	n	−.431
22	n	−1.20
31	n	.456
35	p	−.889
39	n	−1.30
42	n	−1.09
47	n	−.968
56	n	−.878
57	n	−1.03
72	n	−1.40
76	n	−1.26
80	n	−1.16
86	n	−.399
87	n	−1.15
103	p	−.988
105	n	.128
118	n	−.941
144	e	−.787
148	p	−1.10
159	n	−1.50
166	p	−1.20
170	n	−.316
184	e	−.978
185	e	−1.40
194	e	−.291
196	n	−.368
197	n	−.388
204	n	−.878
205	n	−1.05
214	n	−1.02
226	n	−.889
228	n	−1.09
231	n	−.958
249	n	−1.07
256	e	−.893
270	p	.705
271	p	−.912
279	n	−1.26
284	p	−.904
289	n	−1.05
297	n	−1.30
298	n	.503
299	e	−.908
301	p	−1.14
304	p	−.0719
315	n	−.995
324	e	−.961
328	p	−.981
337	n	−.937
340	n	−.515
342	e	−.968
345	e	−1.18
346	p	−.878
350	n	−1.10
351	n	−1.06
363	n	−1.09
388	n	−1.17
396	n	−.901
405	n	−.908
409	e	.201
412	n	−.544
414	e	−1.03
415	n	−.318
428	p	−1.01
438	n	−1.10
450	n	−.912
455	e	−1.03
462	p	.933
465	n	−.268
467	n	−1.32
471	e	−1.01
472	e	1.35

NOTE.—Let A_1 be the derived allele, A_2 be the ancestral allele, and $w(x)$ be the fitness of individuals with genotype x. When $\hat{\gamma} < 0$, $h = 0.5$ implies that $w(A_1A_1) < w(A_1A_2) < w(A_2A_2)$. When $\hat{\gamma} > 0$, $h = 0.5$ implies that $w(A_1A_1) > w(A_1A_2) > w(A_2A_2)$.

[a] As defined by Bejerano et al.[6] Type "e" suggests the element overlaps the mRNA of a known human protein-coding gene (including the UTR regions). Type "n" indicates there is no evidence of transcription of this element from any matching EST or mRNA from any species. Type "p" refers to elements where evidence of transcription is inconclusive.

[b] Scaled s by effective population size.

Web Resources

The URLs for data presented herein are as follows:

BLAT, http://genome.ucsc.edu/cgi-bin/hgBlat
dbSNP, http://www.ncbi.nlm.nih.gov/SNP/
GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for genes in table A3)
Sequenom Inc., http://www.sequenom.com/
The Jackson Laboratory, http://www.jax.org/
UCSC Genome Browser, http://genome.ucsc.edu/

References

1. Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D (2003) The share of human genomic DNA under selection estimated from human-mouse genomic alignments. Cold Spring Harb Symp Quant Biol 68:245–254
2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
3. Bejerano G, Haussler D, Blanchette M (2004) Into the heart of darkness: large-scale clustering of human non-coding DNA. Bioinformatics Suppl 1 20:I40–I48

4. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, et al (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 420: 578–582
5. Margulies EH, Blanchette M, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13:2507–2518
6. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325
7. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, et al (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38:223–227
8. Kryukov GV, Schmidt S, Sunyaev S (2005) Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet 14:2221–2229
9. Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ (2005) Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res 15:1373–1378
10. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12:996–1006
11. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311
12. Jurinke C, van den Boom D, Cantor CR, Koster H (2002) The use of MassARRAY technology for high throughput genotyping. Adv Biochem Eng Biotechnol 77:57–74
13. Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data. Sinauer Associates, Sunderland, MA
14. Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372
15. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76: 887–893
16. Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, United Kingdom
17. Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713–719
18. Takahata N (1986) An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet Res 48:187–190
19. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304
20. Hardison RC (2003) Comparative genomics. PLoS Biol 1:E58
21. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132:1161–1176
22. Williamson S, Fledel-Alon A, Bustamante CD (2004) Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. Genetics 168: 463–475
23. Kim Y (2004) Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. Mol Biol Evol 21:286–294
24. Lu J, Wu CI (2005) Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci USA 102:4063–4067
25. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664
26. Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of ancient human populations. Curr Anthropol 34:483–496
27. Goldberg TL, Ruvolo M (1997) The geographic apportionment of mitochondrial genetic diversity in east African chimpanzees, Pan troglodytes schweinfurthii. Mol Biol Evol 14: 976–984
28. Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2:8
29. Korf I, Flicek P, Duan D, Brent MR (2001) Integrating genomic homology into gene structure prediction. Bioinformatics Suppl 17:S140–S148
30. Wang T, Stormo GD (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 19:2369–2380
31. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90
32. Eddy SR (2005) A model of the statistical power of comparative genome sequence analysis. PLoS Biol 3:e10
33. Roff D (2000) The evolution of the G matrix: selection or drift? Heredity 84:135–142