Human Molecular Genetics, 2005, Vol. 14, No. 15 ***2221–2229***
doi:10.1093/hmg/ddi226
Advance Access published on June 30, 2005

Small fitness effect of mutations in highly conserved non-coding regions

Gregory V Kryukov[1,2], Steffen Schmidt[1,2] and Shamil Sunyaev[1,2]*

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School and [2]Harvard-M.I.T. Division of Health Science and Technology, Harvard Medical School New Research Building, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Received February 9, 2005; Revised May 20, 2005; Accepted June 22, 2005

Comparison of human and mouse genomes has revealed that many non-coding regions have levels of sequence conservation similar to protein-coding genes. These regions have attracted a lot of attention as potentially functional genomic sequences. However, little is known about the effect mutations in these conserved non-coding regions have on fitness and how many of them are present in the human genome as deleterious polymorphisms. To gain insight into the selective constraints imposed on conserved non-coding and protein-coding regions, we compared substitution rates in primate and rodent lineages and analyzed the density and allele frequencies of human polymorphism. Genomic regions conserved between primate and rodent groups show higher relative conservation within rodents than within primates. Thus, our analysis indicates a genome-wide relaxation of selective constraint in the primate lineage, which most likely resulted from a smaller effective population size. We found that this relaxation is much more profound in conserved non-coding regions than in protein-coding regions, and that mutations at a large proportion of sites in conserved non-coding regions are associated with very small fitness effect. Data on human polymorphism are also consistent with very weak selection in conserved non-coding regions. This staggering enrichment in sites at the borderline of neutrality can be explained by assuming an important role for synergistic epistasis in the evolution of non-coding regions. Our results suggest that most individual mutations in conserved non-coding regions are only slightly deleterious but are numerous and may have a significant cumulative impact on fitness.

INTRODUCTION

Comparative analysis of mammalian genomes has detected numerous conserved non-coding regions (1–5). The degree of similarity of human and mouse DNA sequences in these regions is frequently higher than that in protein-coding genes (3,5). This high level of sequence conservation suggests that many mutations in these regions are rejected by natural selection and, therefore, are not phenotypically neutral. However, the extent to which mutations in different regions ultimately affect molecular function and organism's fitness remains an open question. Currently, it is not technologically feasible to study directly the functional effects of mutations in conserved regions on a genome-wide scale. However, we can study the effect of nucleotide substitutions on function indirectly by analyzing the associated forces of natural selection (6).

It has been suggested that the degree of sequence conservation can be used to guide the search for phenotypically important DNA variants, especially in non-coding regions, where other known sources of information might be scarce (7). The reasoning behind this idea is straightforward and intuitively appealing; the higher the conservation of the region, the more detrimental, on average, will be the mutations in this region. However, the relationship between the strength of natural selection and the sequence conservation is truly complex for two reasons. First, a relative loss in fitness associated with mutation of a particular nucleotide affects the probability of this nucleotide being conserved in highly non-linear manner (8). Secondly, mutations of individual nucleotides in a region can have effects on fitness ranging from neutral to extremely deleterious. As a result, an observed overall sequence conservation is an integral value generated by a mixture of sites that is potentially very heterogeneous. Thus, two hypothetical regions with distinct proportions of absolutely crucial, mildly important and neutral sites may have identical sequence conservation.

*To whom correspondence should be addressed. Tel: +1 6175254735; Fax: +1 6175254705; Email: ssunyaev@rics.bwh.harvard.edu

To gain insight into the fine structure of evolutionary forces operating on human genome, we need to venture beyond the analysis of primate/rodent conservation. The availability of two pairs of closely related genomes [human/chimpanzee (9,10) and mouse/rat (1,11)], along with the extensive data on human polymorphisms, allowed us to systematically study and compare the selective constraints in primate versus rodent lineages and in 'conserved protein-coding' versus 'conserved non-coding' regions. Our first major observation from this analysis was that the selective constraints are much stricter in rodents. The proportion of fixed mutations at regions conserved in primates and rodents was lower in the rodent lineage than that in the primate lineage. The higher effective size of the rodent population is translated into a more efficient purifying selection process which, in turn, leads to the accumulation of fewer slightly deleterious substitutions. The difference in the rates of accumulation of substitutions in primate and rodent lineages is particularly apparent in conserved non-coding regions, where primates have accumulated significantly more slightly deleterious mutations than rodents have. Although this effect was predicted by classical population genetics (8), the magnitude evident from a genome-wide analysis is intriguing. This profound relaxation of selective constraints in the primate lineage relative to the mouse lineage can be explained only if many mutations in these regions are associated with small, almost neutral, effects on fitness. If the selection coefficients of many mutations in a region are 'concentrated' at the border between random genetic drift and natural selection, a several fold decrease of the effective population size and a corresponding elevation in genetic drift will lead to an observed sharp increase in rates of accumulation of substitutions.

To gain a better understanding of the evolutionary mechanisms underlying the observed results, we estimated the distributions of selection coefficients associated with nucleotide substitutions in conserved protein-coding, intronic and intergenic regions. For all categories of regions, we analyzed a range of parameters: (i) the number of substitutions per site accumulated between human and chimpanzee DNA sequences (K_{hc}); (ii) the number of substitutions per site accumulated between mouse and rat DNA sequences (K_{mr}); (iii) the nucleotide diversity within the human population (π) and (iv) the fraction of polymorphic variants in the human population with the frequency of a derived allele $<5\%$ ($F_{0.05}$). The theoretical dependence of these parameters on the selection coefficient has been well-studied in population genetics (8,12,13). We computationally generated a large spectrum of possible probability density functions for selection coefficient and determined which of them produced parameter values similar to the observed values. The analysis of estimated probability density functions revealed their full correspondence with a more qualitative interpretation of our data discussed earlier.

As a concluding step of our analysis, we propose a model of evolution of non-coding regions that explains this staggering enrichment in sites at the borderline of neutrality. The key assumption of the model is the presence of synergistic epistasis (14) between individual substitutions, when each new mutation has a greater impact on their accumulation. Under such a model, the observed distribution of selection coefficients emerges as an equilibrium state, which resulted from the interplay between degrading mutational pressure and natural selection.

RESULTS AND DISCUSSION

Divergence and polymorphism in conserved genomic regions

We determined human–chimpanzee divergence (K_{hc}) and mouse–rat divergence (K_{mr}) in conserved protein-coding, intronic and intergenic regions. In protein-coding regions, the values of K_{hr} and K_{mr} were determined separately for three categories of sites: (i) non-degenerate codon positions at which any mutation is causing a substitution in amino acid sequence; (ii) 4-fold degenerate codon positions at which all possible substitutions are synonymous and (iii) all nucleotide positions combined (including 2-fold and 3-fold degenerate sites). We also estimated the values for K_{hc} and K_{mr} in putative neutrally evolving genomic regions. Data on K_{hc} and K_{mr} are presented in Figure 1 (A and B, respectively). As CpG dinucleotides prone to methylation are known to mutate in mammals at extremely high rate which exceeds genome-wide average by an order of magnitude, inclusion of these sites may result in biased estimates of divergence and polymorphism rates. To avoid various biases introduced by hypermutable CpG nucleotides, we exclude all nucleotide positions preceded by C or followed by G from all divergence and nucleotide diversity calculations. However, we note that inclusion of CpG sites does not change the results and conclusions of our analysis.

To compare the data on divergence in different lineages directly, the obtained values need to be normalized to the neutral level of divergence observed in the corresponding lineage in the absence of selective pressure. For putative neutrally evolving regions, we now used accurately filtered non-coding non-repetitive genomic regions remote from splice sites and putative transcription initiation and termination sites. Predicted putative exons (including suboptimal) were also excluded.

For all regions, we found that reduction in divergence relative to the neutral level was more pronounced in the rodent lineage than that in the primate lineage ($K_{mr}/K^0_{mr} < K_{hc}/K^0_{hc}$). The difference in the number of fixed mutations normalized to the neutral divergence between human and mouse lineages was especially profound in conserved non-coding sequences—a 2.1-fold in intergenic and a 1.7-fold in intronic regions. In contrast, only a 12% difference was observed in protein-coding regions, although we detected a 1.9-fold difference, if only non-degenerate sites were considered. The reduction of selective constraint for non-synonymous substitutions in the human lineage was earlier studied by Eyre-Walker *et al.* (15), and the very strong widespread relaxation of selection in promotor regions was recently reported by Keightley *et al.* (16).

Note that the levels of divergence between mouse and rat DNA sequences in conserved coding, intronic and intergenic regions, which we have selected, are similar. These regions were selected using the same conservation threshold between primates and rodents. Because rodent evolution is

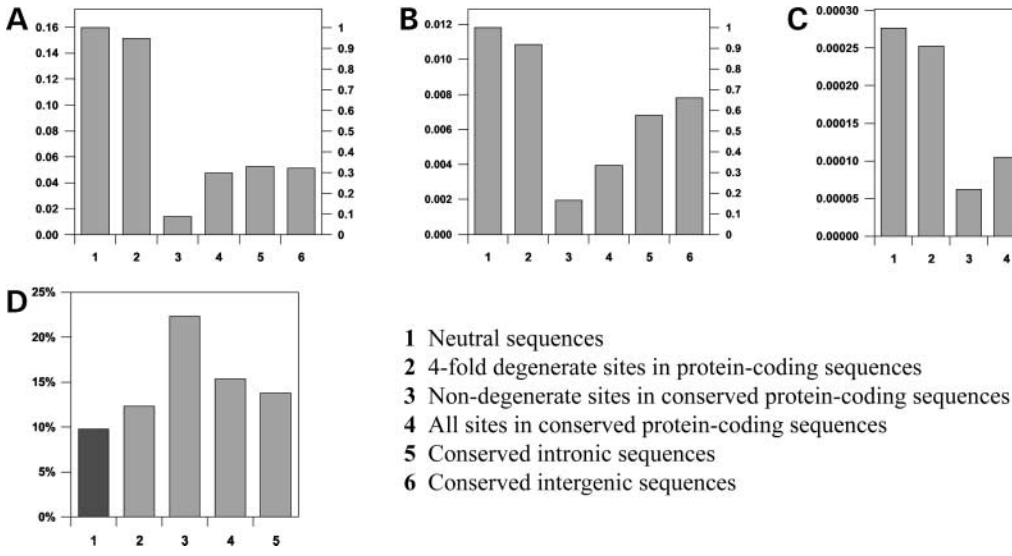

Figure 1. Selective constraints in human and mouse lineages: (**A**) number of substitutions (per nucleotide) accumulated between mouse and rat (K_{mr}/K^0_{mr}); (**B**) number of substitutions (per nucleotide) accumulated between human and chimpanzee (K_{hc}/K^0_{hc}); (**C**) nucleotide diversity in the human population (π/π^0) and (**D**) fraction of polymorphic variants in the human population with the frequency of a derived allele $<5\%$ ($F_{0.05}$). A calculated theoretical value corresponding to the absence of selection is shown as a genome average for $F_{0.05}$. In Figures A, B and C the left scale shows absolute values and the right shows values normalized to the neutral level.

generally faster than primate evolution in terms of the number of accumulated substitutions, the conservation between primates and rodents is more characteristic for the rodent lineage. It is also possible that the effective population size of ancestral populations in both lineages was closer to rodent population size than to primate population size.

The observed increase in the rate of substitution accumulation in the human lineage in conserved non-coding genomic regions can be explained in two ways: as a massive, genome-wide increase in positive selection acting on functional non-coding regions of the human genome or as a relaxation of purifying selection acting on these regions. We can distinguish between these two possibilities by analyzing the polymorphism in the human population. Under positive selection, the decrease in divergences relative to the neutral level is expected to be more pronounced than the corresponding decrease in nucleotide diversity, for the same genomic region in the same population (13,17). In contrast, under relaxed selective constraints, the relative decrease in nucleotide diversity should be lower than the relative decrease in divergence. We calculated the nucleotide diversity (π) in conserved protein coding, conserved non-coding and putatively neutrally evolving regions of the human genome. A set of almost a million polymorphic sites identified as differences between Celera individual A and NCBI human genome assembly was used for such calculation. Other publicly available data sets of human single nucleotide polymorphisms (SNPs) obtained by comparison of shotgun reads give qualitatively highly similar results, although numerical values of estimates vary.

Calculated values of π in different genomic regions are shown in Figure 1C. The relative decrease in nucleotide diversity when compared with the neutral level was slightly lower than the relative decrease in the number of substitutions accumulated in the human lineage, indicating the transient presence of deleterious mutations in the human genome in the form of polymorphisms (18,19). The number of such mutations appears to be large, as the difference between π/π^0 and K_{hc}/K^0_{hc} is apparent from a genome-scale analysis. A clear trend is evident in the conserved non-coding regions: $K_{mr}/K^0_{mr} < K_{hc}/K^0_{hc} < \pi/\pi^0$. On the basis of these data, we proposed that the observed increase in the rate of accumulation of nucleotide substitutions in conserved non-coding regions in the primate lineage is a product of the reduction in the effective population size that resulted in increased random genetic drift. However, the magnitude of the increase can be explained only if we assume that a very high proportion of sites in non-coding regions was subject to a weak selection on the borderline of neutrality.

The theoretical dependence of K_{mr}/K^0_{mr}, K_{hc}/K^0_{hc} and π/π^0 on selection coefficient s is presented in Figure 2. The decline in the divergence occurs within a relatively narrow range of values of selection coefficients and occurs at higher values of s for smaller populations. Sites at which substitutions are associated with selection coefficients close to 10^{-5} mutate at almost a neutral rate in the primate lineage but are highly conserved in the rodent lineage because of its larger population size. We propose that many sites in conserved non-coding regions of mammalian genomes are subject to such weak selection. Under such model, even a moderate increase in the strength of random genetic drift in the human lineage would lead to the conservation of a large proportion of nucleotides in non-coding sequences in rodents, but not in humans.

Not only are the number of polymorphic variants strongly affected by the strength of natural selection acting on a genomic region but so are the frequencies of these variants in a population. Strict selective constraints lead to a decrease in the average frequency of new variants, which are usually deleterious. We calculated the fraction of SNPs with a derived allele frequency $<5\%$ ($F_{0.05}$) in conserved protein-coding and

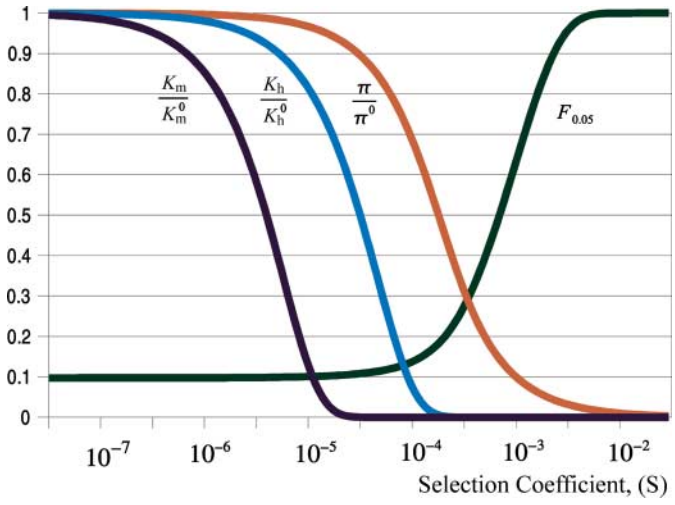

Figure 2. Theoretical dependence on selection coefficient for number of substitutions between mouse and rat (normalized to the neutral level) [dark blue line (K_{mr}/K_{mr}^0)]; number of substitutions between human and chimpanzee (normalized to the neutral level) [light blue line (K_{hc}/K_{hc}^0)]; nucleotide diversity in the human population (normalized to the neutral level) [red line (π/π^0)] and fraction of polymorphic variants in the human population with the frequency of derived allele $<5\%$ [green line ($F_{0.05}$)].

non-coding regions of the human genome. We used $F_{0.05}$ as a measure of the excess of rare-derived allelic variants. The calculated values $F_{0.05}$ in different genomic regions are shown in Figure 1D and their theoretical dependence on the strength of selective pressure is shown in Figure 2. The excess of new rare alleles was the highest for non-degenerate sites in protein-coding sequences, followed by the value for all sites in conserved protein-coding regions and then by the intronic and intergenic regions. This observation is in accord with our current beliefs that non-synonymous substitutions are the most deleterious form of single nucleotide mutations.

Distribution of selection coefficients

To characterize the observed differences in selective pressure acting on the conserved protein and non-protein coding regions in a more rigorous, quantitative manner, we estimated the distribution of selection coefficients in these regions using a computational 'exhaustive search' method. We generated a large spectrum of possible probability density functions of selection coefficients and determined which of them produced values of K_{hc}, K_{mr} and $F_{0.05}$ similar to the observed values. In previous reports, the probability density function of selection coefficients for protein-coding regions was modeled by a gamma function (20). However, there is no reason to believe that the same function can be applied to non-coding regions. For maximal generality, we modeled the probability density function describing the distribution of selection coefficients by a histogram.

The distributions that provided the best fit for the observed values of K_{hc}, K_{mr}, π and $F_{0.05}$ parameters in conserved protein-coding, intronic and intergenic regions are shown in Figure 3. Our results indicate that 80% of non-degenerate sites are associated with selection coefficients higher than $10^{-3.5}$, indicating that the majority of non-synonymous substitutions have significant deleterious effects. In contrast to non-degenerate sites, entire conserved protein-coding regions contain a significant proportion of effectively neutral sites, mutations at which selection coefficients are less than $10^{-6.5}$. The proportion of such sites roughly corresponds to the proportion of synonymous sites in protein-coding sequences, mutations at which are considered to be mostly neutral in mammals. Our analysis suggests that mutations in $\sim$55% of sites in conserved intergenic regions are associated with small, but non-zero values of selection coefficients more than $10^{-5.5}$ but less than $10^{-3.5}$. In conserved intronic sequences, $\sim$35% of sites fall into this category.

Proposed role of synergistic epistasis in the evolution of non-coding regions

We found that mutations in a large proportion of sites in conserved non-coding regions have selection coefficient values approaching $1/N_e$ (where N_e is an effective population size). To understand this interesting feature, one should consider factors that affect the accumulation of substitutions at such mildly important sites. The fate of a slightly deleterious mutation is determined by the interplay of selection and random genetic drift. The strength of natural selection becomes smaller than the effect of random genetic drift, when the product of effective population size and selection coefficient is close to 1. Such a 'neutrality boundary' for the mouse population corresponds approximately to selection coefficients equal to 10^{-5}. For the human population, which has a smaller effective size, this boundary corresponds to selection coefficients equal to 10^{-4}.

The calculated distributions of selection coefficients (Fig. 3C and D) indicate that the conserved non-coding regions contain a high proportion of sites with selection coefficients, which lie between neutrality boundaries for the mouse and human populations. The phenomenon of synergistic epistasis offers an excellent explanation for the observed effect of the enrichment of conserved non-coding sequences in sites on the border of neutrality [a similar reasoning was initially proposed by Akashi (21,22) in the analysis of weak selection on synonymous substitutions]. The assumption of synergistic interactions between sites implies that an individual mutation has a greater effect on an organism that is already loaded with mutations. The higher the number of deleterious mutations in the region, the greater the impact on fitness (higher selective coefficient) of a new mutation.

Consider a functional genomic region composed of nucleotides with their small individual effects contributing to the absolute fitness. Mutation at each site is more likely to be deleterious than beneficial and, on average, to result in the divergence of a sequence from the optimum by mutational pressure. With each new fixed mutation, all remaining non-mutated sites become slightly more important, as follows from our assumption of synergistic interactions. The equilibrium state of this process corresponds to the point at which the selection coefficients of the majority of sites are just high enough to prohibit the fixation of new mutations. The probability of fixation of a new allele becomes close to zero when the selection coefficients exceed $1/N_e$. The process of evolution

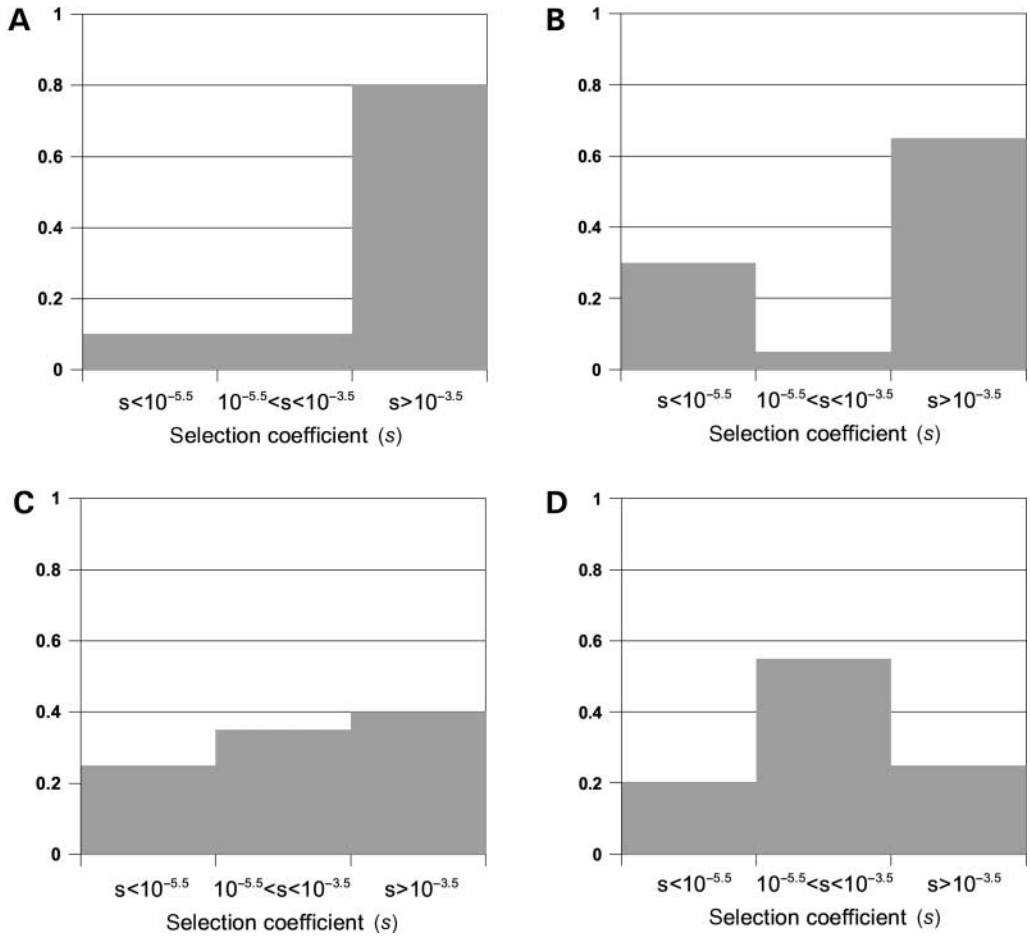

Figure 3. Estimated distributions of selection coefficients in conserved genomic regions. (**A**) Non-degenerate sites in protein-coding sequences; (**B**) all sites in protein-coding sequences; (**C**) intronic sequences and (**D**) intergenic sequences.

of a region can be considered as its traveling between distinct 'states' with various numbers of sites that differ from the region's optimal sequence. In the model we considered, the fitness depends only on the total number of nucleotides that match the optimal sequence. The probability of transition between different 'states' is affected by selective pressure and random genetic drift. The exact solution for the probability distribution of site' selection coefficients, which corresponds to equilibrium steady state of the described process, can be obtained by applying the mathematical framework of 'birth–death processes' that we used (see Data, formulas and methods) (23). The same solution can be obtained within framework of statistical physics (24). To illustrate the model, we considered a sample function that describes the fitness dependence on the number of nucleotides that differ from a region's optimal sequence. Such a function was chosen to satisfy two criteria. First, to reflect a behavior characteristic of synergistic epistasis, a decline in fitness with an accumulation of mutations should be slower than the exponential. Secondly, the selection coefficient associated with a first substitution in an optimal sequence should be lower than the random genetic drift boundary; otherwise, the region will maintain 100% conservation. The sample function and calculated equilibrium probability density of corresponding selection coefficients are shown in Figure 4 (A and B, respectively). Indeed, we can observe a peak at selection coefficients close to the random genetic drift boundary.

CONCLUSION

One of the major advantages of comparative genomics is in the identification of highly conserved genomic regions (7). Sequence conservation was proposed as particularly useful for the detection of potentially functional regions that can harbor mutations and polymorphic variants affecting phenotypes (7). Numerous studies are aimed at identifying frequent polymorphic variants of large effect on specific phenotypes (25). It was hypothesized that, unlike Mendelian diseases, which are caused predominantly by mutations in protein-coding regions (26), complex phenotypes may also be inherited through sequence changes in conserved non-protein-coding regions (6,25). Genetic variation in *cis*-acting loci has been shown to contribute to variation in gene-expression levels (27–29), although, to date, the success in identification of the *cis*-acting regulatory single nucleotide variants has been limited (30).

However, our analysis shows that an average mutation in a highly conserved non-coding region is much less likely than a

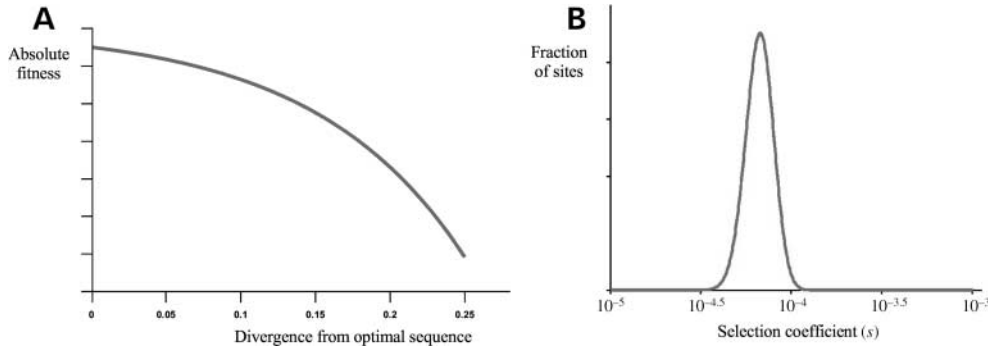

Figure 4. Modeled distribution of selection coefficients of sites. (**A**) Sample function describing fitness dependence on a region's divergence from its optimal sequence in a manner consistent with synergistic epistasis model; (**B**) corresponding distribution of selection coefficients.

mutation in a protein-coding region to have a large effect on fitness, and presumably on phenotype. Therefore, sequence conservation is not a very reliable guide when searching for polymoprhic variants of large effect. In contrast, polymorphic variants of small effect are numerous in conserved non-coding regions. Their cumulative effect may be substantial, and inheritance of some phenotypes may be explained by a large number of simultaneously acting non-coding variants.

DATA, FORMULAS AND METHODS

Selection of annotated protein-coding, intronic and intergenic regions

Multiple genome alignments (31) and annotation tracks were obtained from University of California, Santa Cruz (UCSC), Genome Bioinformatics FTP site (32). Human genome assembly hg17, chimpanzee genome assembly PanTrot1, mouse genome assembly mm5 and rat genome assembly rn3 were used. Protein-coding regions are defined as regions that are annotated as protein coding in the refGene (33) track and have no conflicting annotations in either the refGene track or the knownGene track (based on proteins from SWISS-PROT, TrEMBL and TrEMBL-NEW (34) and their corresponding mRNAs from GenBank). Intronic sequences are defined as regions that are annotated as intronic in the refGene track, have no conflicting annotations in the refGene and knownGene tracks, and were not predicted to be protein coding by GeneID (35) and GENSCAN (36) programs (tracks geneid, genscan and genscanSubopt). Intergenic regions are defined as regions that are not annotated as genes in the RefGene, KnowGenes, mrna, GeneID, GeneScan and GeneScanSubopt tracks, do not correspond with any expression sequence tag (EST) and are not located within gene boundaries predicted from EST cluster analysis (rnacluster track). Fifty nucleotides adjacent to exon/intron boundaries were excluded from intronic and intergenic sequences. The neutral divergence and polymorphism rate were estimated on the basis of combined intronic and intergenic sequences, which are remote from splice sites (for at least 0.5 kb) and putative transcription initiation and termination sites (for at least 2 kb). Sequences corresponding to repetitive elements predicted by RepeatMasker (Smit and Green, unpublished data) were excluded from all types of regions.

Selection of conserved regions

Selection of conserved regions on the basis of human/mouse nucleotide identity restrict a number of bases with substitution in human lineage after divergence with chimpanzee and in mouse lineage after divergence with rat. This would lead to an underestimation of K_{hc} and K_{mr}. The bias can be avoided if different nucleotides are used in the process of choosing conserved regions and in calculating divergence. Regions of primate/rodent conservation were identified as 50 nucleotide windows that have human/mouse sequence identity $>90\%$ at odd (not divisible by two) nucleotide positions.

Conserved genomic regions selected using the aforementioned procedure mostly coincide with conserved regions selected based on simple sequence identity. Further, the analysis of conserved regions selected on the basis of simple sequence identity produces qualitatively similar results.

Calculation of divergence

The numbers of substitutions between human and chimpanzee and between mouse and rat were calculated at nucleotide positions that passed following filters. First, to avoid mistakes introduced by ambiguous alignment, nucleotide positions that are adjacent to gaps or unidentified nucleotides were excluded from the analysis. Secondly, to exclude the effect of hypermutable CpG dinucleotides, positions that were preceded by C or followed by G in either of two closely related genomes were excluded from the analysis. Thirdly, to avoid bias introduced by the procedure of choosing conserved regions, nucleotides at odd positions that were used to identify these regions were excluded from further analysis. The divergence was calculated from the observed number of accumulated substitutions per site using the Jukes–Cantor correction (37).

Analysis of polymorphism

As a best possible proxy for nucleotide diversity, we calculated a polymorphism density using 917853 random SNPs

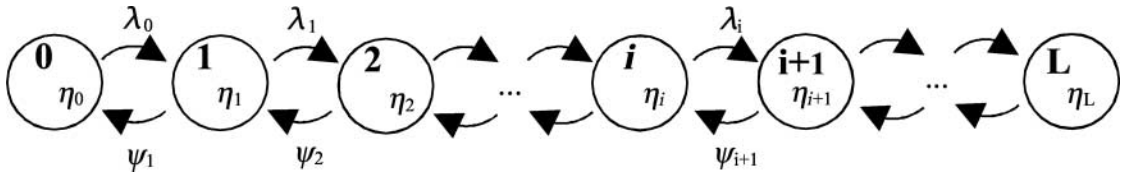

Figure 5. Representation of sequence evolution in the framework of birth–death processes. L is the sequence length, η_i the probability of a state with i 'bad' nucleotides, λ_i the probability of the region transition from a state with i 'bad' nucleotides into a state with $i+1$ 'bad' nucleotides and ψ_{i+1} the probability of the region transition from a state with $i+1$ 'bad' nucleotides into a state with i 'bad' nucleotides.

detected by alignment of sequencing reads from individual A of Celera Human Genome sequencing project (10) to the NCBI human genome assembly (9) (build34). Density of polymorphisms identified by comparison of two genomic sequences should be, by definition, directly proportional to nucleotide diversity. Individual A was chosen as one that has the highest number of detected SNPs (presumably due to best genome coverage). SNP set was obtained from dbSNP (build 123, method_id = 1550) (38). The fractions of new rare alleles were estimated using the data on allele frequencies generated by 'A database of Japanese Single Nucleotide Polymorphisms' (JSNP) project (39). The ancestral allele was determined from the chimpanzee sequence.

Estimation of selection coefficients

Possible distributions of selection coefficients (s) were modeled by a 10-column histogram that contained bins corresponding to neutral sites ($s < 10^{-6.5}$), sites under very strong selection ($s > 10^{-2.5}$) and eight intermediate categories of sites with selection coefficients in the range of $10^{-6.5}$–$10^{-2.5}$. Proportions of sites were allocated to each histogram category in 5% minimal 'blocks'. All possible histograms in this grid were exhaustively searched. Histograms of selection coefficients best reproducing the observed data on polymorphism and divergence were chosen. To guarantee the stability of obtained histograms with respect to small deviation in the observed values, and for data representation, 10 categories were joined into three: neutral (sites with a selection coefficient lower than $10^{-5.5}$), slightly deleterious (sites with a selection coefficient higher than $10^{-5.5}$ but lower than $10^{-3.5}$) and deleterious (sites with a selection coefficient higher than $10^{-3.5}$).

Theoretical values for the number of accumulated substitutions corresponding to a particular selection coefficient, in assumption of no dominance, were calculated using the following formula (8,13):

$$K = K_0 2N_e \frac{(1 - e^{-2s})}{(1 - e^{-4N_e s})} \quad (1)$$

where K is the divergence, K_0 the neutral divergence in the absence of natural selection, s the selection coefficient and N_e the population effective size.

The effective sizes of the mouse and human populations were considered to be 85 000 (40) and 10 000 (41), respectively. Theoretical values for nucleotide diversity for a particular selection coefficient were calculated as follows:

$$\pi = \pi_0 \frac{2N_e s + e^{-2N_e s} - 1}{N_e s(1 - e^{-2N_e s})} \quad (2)$$

where π is the nucleotide diversity, π_0 the neutral nucleotide diversity in the absence of natural selection and N_e the effective population size.

A theoretical value for the fraction of rare alleles was calculated with the following formula (8,13):

$$F_{0.05} = \frac{\left[\int_0^{0.05} (e^{-2N_e s(1-x)} - 1/x(1-x)(e^{-2N_e s} - 1))(1 - x^m - (1-x)^m)dx\right]}{\left[\int_0^{1} (e^{-2N_e s(1-x)} - 1/x(1-x)(e^{-2N_e s} - 1))(1 - x^m - (1-x)^m)dx\right]} \quad (3)$$

where m is the number of sequences used for SNP detection. We assume here that after SNP detection in a set of m individuals, allele frequencies were estimated by genotyping in a very large sample of individuals.

Parameter m was set to 2, because using this value produces the best fit for the practically linear tail of distribution of new allele frequencies observed for the entire JSNP data set. To measure the similarity of the theoretical values of K_h, K_m, π and $F_{0.05}$ to the observed ones, the following empirical measure of dissimilarity was used:

$$\Delta = \frac{(K_m^{\text{theoretical}} - K_m^{\text{observed}})^2}{K_m^{\text{theoretical}}} + \frac{(K_h^{\text{theoretical}} - K_h^{\text{observed}})^2}{K_h^{\text{theoretical}}} + \frac{(\pi^{\text{theoretical}} - \pi^{\text{observed}})^2}{\pi^{\text{theoretical}}} + \frac{1}{10}\frac{(F_{0.05}^{\text{theoretical}} - F_{0.05}^{\text{observed}})^2}{F_{0.05}^{\text{theoretical}}} \quad (4)$$

For Δ calculation, we assumed that $K_m = K_{mr}/2$ and $K_h = K_{hc}/2$.

There are very few SNPs in JSNP data set with known frequencies located in intergenic regions. Because of that only K_m, K_h and π terms were used to calculate values of Δ in conserved intergenic regions.

Synergistic epistasis through the framework of birth–death processes

Evolution of a genomic region can be considered within a mathematical framework of 'birth–death' processes. A region comprising L nucleotides can exist in $L+1$ distinct 'states' corresponding to a number of deleterious substitutions it contains (i.e. nucleotides that differ from some optimal consensus sequence) (Fig. 5). For simplicity, consider that, at each position, only one of four nucleotides is favorable (a 'good' nucleotide) and the other three are deleterious ('bad' nucleotides). We are interested in the equilibrium distribution

across all possible states (η_i at equilibrium). This distribution can be considered as either the probabilities of one particular region being at the particular state or the proportion of many genomic regions that are at some particular states.

Under equilibrium, net flow at each state is equal to zero:

$$\eta_i \lambda_i = \eta_{i+1} \psi_{i+1} \tag{5}$$

The recursive formula for an equilibrium distribution can be obtained:

$$\eta_{i+1} = \frac{\lambda_i}{\Psi_{i+1}} \eta_i \tag{6}$$

where λ_i is the probability of the region transition from a state with i 'bad' nucleotides into a state with $i+1$ 'bad' nucleotides and ψ_{i+1} is the probability of the region transition from a state with $i+1$ 'bad' nucleotides into a state with i'bad' nucleotides.

Transition probabilities from state to state λ_i and ψ_{I+1} can be calculated as follows:

$$\lambda_i = \mu(L-i)\frac{(1-e^{2s})}{(1-e^{4Ns})} \tag{7}$$

$$\Psi_{i+1} = \mu\frac{1}{3}(i+1)\frac{(1-e^{-2s})}{(1-e^{-4Ns})} \tag{8}$$

where μ is the mutation rate, 1/3 is the probability of a mutation to be beneficial (to convert a 'bad' nucleotide into a 'good' one), N the effective population size and s the selection coefficient with a new mutation. s is calculated as follows:

$$s = 1 - \frac{F_{i+1}}{F_i} \tag{9}$$

where si is the selection coefficient associated with mutations in the regions that contains i 'bad' nucleotides and F_i is the fitness associated with a region that contains i 'bad' nucleotides.

To incorporate interactions between deleterious mutations characteristic of synergistic epistasis, the following formula for fitness depending on the number for deleterious mutations ('bad' nucleotides) was used:

$$F_i = \frac{1}{1 + 10^{-4}\ e^{10(i/L)}} \tag{10}$$

η_0 can be obtained from consideration of normalization that the sum of all η_i should be equal to 1.

ACKNOWLEDGEMENTS

The authors would like to thank Drs Alexey Kondrashov and Leonid Mirny for helpful discussions on the manuscript and Dr Ivan Adzhubey for technical support. This work was funded by the Genome Canada Foundation and National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54LM008748. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics.

Conflict of Interest statement. None declared.

REFERENCES

1. Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
2. Kondrashov, A.S. and Shabalina, S.A. (2002) Classification of common conserved sequences in mammalian intergenic regions. *Hum. Mol. Genet.*, **11**, 669–674.
3. Dermitzakis, E.T., Kirkness, E., Schwarz, S., Birney, E., Reymond, A. and Antonarakis, S.E. (2004) Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.*, **14**, 852–859.
4. Bejerano, G., Haussler, D. and Blanchette, M. (2004) Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics*, **20**, i40–i48.
5. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
6. Keightley, P.D. and Gaffney, D.J. (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl Acad. Sci. USA*, **100**, 13402–13406.
7. Boffelli, D., Nobrega, M.A. and Rubin, E.M. (2004) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.*, **5**, 456–465.
8. Kimura, M. (1994) *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers.* The University of Chicago Press, Chicago, IL, pp. 69–72.
9. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
10. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
11. Rat Genome Sequencing Project Consortium. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
12. Haldane, J.B.S. (1924) A mathematical theory of natural and artificial selection. *Proc. Camb. Philos. Soc.*, **23**, 838–844.
13. Sawyer, S.A. and Hartl, D.L. (1992) Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.
14. West, S.A., Peters, A.D. and Barton, N.H. (1998) Testing for epistasis between deleterious mutations. *Genetics*, **149**, 435–444.
15. Eyre-Walker, A., Keightley, P.D., Smith, N.G. and Gaffney, D. (2002) Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.*, **19**, 2142–2149.
16. Keightley, P.D., Lercher, M.J. and Eyre-Walker, A. (2005) Evidence for widespread degradation of gene control regions in hominid genomes. *PloS Biol.*, **3**, 282–288.
17. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila. Nature*, **351**, 652–654.
18. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., III and Kondrashov, A.S. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
19. Fay, J.C., Wyckoff, G.J. and Wu, C.I. (2001) Positive and negative selection on the human genome. *Genetics*, **158**, 1227–1234.
20. Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**, 512–526.
21. Akashi, H. (1995) Quantifying the slightly deleterious mutation model of molecular evolution. *Genetics*, **139**, 1067–1076.
22. Akashi, H. (1996) Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster. Genetics*, **144**, 1297–1307.
23. Jones, P.W. and Smith, P. (2001) *Stochastic Processes: An Introduction.* Arnold Publishers, London, UK, pp. 133–141.
24. Berg, J., Willman, S. and Lässig, M. (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.*, **4**, 42.
25. Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33**, 228–237.
26. Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.

27. Cowles, C.R., Hirschhorn, J.N., Altshuler, D. and Lander E.S. (2002) Detection of regulatory variation in mouse genes. *Nat. Genet.*, **32**, 432–437.
28. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
29. Wittkopp, P.J., Haerum, B.K. and Clark, A.G. (2004) Evolutionary changes in cis and trans gene regulation. *Nature*, **430**, 85–88.
30. Hudson, T.J. (2003) Wanted: regulatory SNPs. *Nat. Genet.*, **33**, 439–440.
31. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D. *et al*. (2004) Aligning multiple genomic sequences with threaded blockset aligner. *Genome Res.* **14**, 708–715.
32. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
33. Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
34. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E, Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
35. Parra, G., Blanco, E. and Guigo, R. (2000) GeneID in *Drosophila. Genome Res.*, **10**, 511–515.
36. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
37. Li, W.H. (1997) *Molecular Evolution.* Sinauer Associates, Sunderland, MA, pp. 59–62.
38. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
39. Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T. and Nakamura, Y. (2002) JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.*, **30**, 158–162.
40. Nachman, M.W. (1997) Patterns of DNA variability at X-linked loci in Mus domesticus. *Genetics*, **147**, 1303–1316.
41. Takahata, N., Satta, Y. and Klein, J. (1995) Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.*, **48**, 198–221.