J Mol Evol (2007) 65:236–248
DOI 10.1007/s00239-007-9008-x

Similarity of Synonymous Substitution Rates Across Mammalian Genomes

Jeffrey H. Chuang · Hao Li

Received: 5 September 2006 / Accepted: 4 April 2007 / Published online: 3 August 2007

Abstract Given that a gene has a high (or low) synonymous substitution rate in one mammalian species, will it also have a high (or low) synonymous substitution rate in another mammalian species? Such similarities in the rate of synonymous substitution can reveal both selective pressures and neutral processes acting on mammalian gene sequences; however, the existence of such an effect has been a matter of disagreement. We resolve whether such synonymous substitution rate similarities exist using 7462 ortholog triplets aligned across rat, mouse, and human, a dataset two orders of magnitude larger than previous studies. We find that a gene's synonymous substitution rate in the rat-mouse branch of the phylogeny is correlated with its rate in the branch connecting human and the rat-mouse ancestor. We confirm this for several different measures of synonymous substitution rate, including corrections for base composition and CpG dinucleotides, and we verify the results in the larger mouse-human-rat-dog phylogeny. This similarity of rates is most apparent for genes in which synonymous sites are well conserved across species, suggesting that a significant component of the effect is due to purifying selection. We observe rate correlations at a resolution as fine as a few hundred kilobases, and the genes with the most similar synonymous substitution rates are enriched for regulatory functions. Genes with above-average substitution rates also exhibit significant, though somewhat weaker, rate correlations, suggesting that some neutral processes may have persisted in the phylogeny as well.

Keywords Synonymous substitution · Repeatability · Neutral mutation · Mammalian genes · Selection · Correlation · Similarity · Gene specificity

Reviewing Editor: Dr. Richard Kliman

Electronic supplementary material The online version of this article (doi: 10.1007/s00239-007-9008-x) contains supplementary material, which is available to authorized users.

J. H. Chuang (✉)
Department of Biology, Boston College, 140 Commonwealth Ave., Chestnut Hill, Massachusetts 02467, USA
e-mail: chuangj@bc.edu

H. Li
Department of Biochemistry and Biophysics, University of California, California, San Francisco 94143, USA

Introduction

An important issue in comparative genomics is the similarity of substitution rates across species, i.e., whether the substitution rate of a sequence in one branch of a phylogeny is correlated with the substitution rate of its orthologous sequence in other branches (Chimpanzee Sequencing Consortium 2005; Hurst and Pal 2001; Smith and Hurst 1998; Smith et al. 2002; Williams and Hurst 2002). This similarity of substitution rates across species has been previously referred to as repeatability (Smith and Hurst 1998), similarity of relative rates (Bulmer et al. 1991), rate constancy (Langley and Fitch 1974), and, in the context of gene sequences, as gene specificity of substitution rates (Mouchiroud et al. 1995; Williams and Hurst 2002). Following the terminology of Smith and Hurst, if a gene has a similar substitution rate in two lineages, we say that the gene's substitution rate is "repeatable."

Such repeatability in substitution rates can indicate both selective pressures and neutral processes. For example, if a sequence has a low substitution rate among several species,

it is common to take this as evidence that the sequence is functional and has been under purifying selection in each species (Boffelli et al. 2003; Hardison 2003; Rat Genome Sequencing Consortium 2004; Xie et al. 2005). This is simply another way of saying that the sequence is conserved across multiple species. On the other hand, similarities in substitution rate can also be informative for understanding the neutral processes that influence mutation and how these processes may change over time. For example, Bulmer et al. (1991) searched for local neutral mutation processes that have persisted across species by comparing synonymous site substitution rates in orthologous mammalian genes, under the assumption that such synonymous sites are neutral. Recently, these selective and neutral perspectives have intersected, as it has become apparent that sites that had previously been presumed to be neutrally evolving and, particularly, mammalian synonymous sites, can also be under selection.

A central characteristic of the molecular evolution of mammalian genomes is that neutral mutation rates vary by location (Chimpanzee Sequencing Consortium 2005; Chuang and Li 2004; Gaffney and Keightley 2005; Hardison et al. 2003; Lercher et al. 2004; Smith et al. 2002; Wolfe et al. 1989). Because synonymous sites in coding sequences do not affect the encoded protein sequence, these sites have often been used to assess the regional neutral processes acting within genomes (Graur and Li 2000). Early studies into the cross-species similarity of mammalian synonymous substitution rates were conducted more than a decade ago by Bulmer et al. (1991) and Mouchiroud et al. (1995), who each analyzed datasets of several tens of genes and concluded that substitution rates were correlated between orthologous genes in different mammalian lineages. Because mammalian synonymous sites were believed to be neutral, this finding was interpreted to indicate that local neutral mutation processes (for both high and low substitution rates) could persist across species. These results were later called into question by Williams and Hurst (2002), who analyzed 116 genes and concluded that the previously observed correlations were artifacts of certain mutation models. These conflicting findings suggest the usefulness of a larger study that could provide more robust conclusions.

More recently, several groups studying selection on mammalian synonymous sites have observed strong multispecies synonymous site conservation in blocks within a number of genes (Chamary and Hurst 2004; Chamary et al. 2006; Hurst and Pal 2001; Schattner and Diekhans 2006; Smith and Hurst 1998; Xing and Lee 2005). It has also been observed that synonymous substitution rates are generally lower than those of pseudogenes (Bustamante et al. 2002) or intergenic sequence (Hellmann et al. 2003). Such studies have provided evidence for purifying selection (reviewed in Chamary et al. 2006), some of which may be explained by functions related to splicing (Fairbrother et al. 2004; Parmley et al. 2006; Xing and Lee 2005), mRNA secondary structure (Chamary and Hurst 2005), or transcription (Kudla et al. 2006). Purifying selection would be expected to lead to repeatable (low) synonymous substitution rates, a prediction incongruous with the earlier studies in which rates were found to not be similar across species. This inconsistency could potentially be resolved through a more complete characterization of genes' synonymous substitution rates and through a dissection of which genes have the most repeatable synonymous substitution rates.

In this article we revisit the issue of repeatability of synonymous substitution rates, using genome-scale datasets to resolve the previous discrepancies. We use data from the mouse, human, and rat genomes, with of the order 70× more genes than in previous studies. First, we clarify whether synonymous substitution rates are repeatable by measuring correlations in substitution rates in two separate lineages. We find that these correlations are significant, and we test the robustness of this finding using several different models for inferring the synonymous substitution rate. We then assess whether selective and/or persistent neutral effects contribute to these correlations. Finally, we extend our results to a set of orthologous genes shared among mouse, human, rat, and dog. This allows us to further validate our results and also characterize how the repeatability of synonymous substitution rates—and hence the selective pressures and neutral processes on each gene's synonymous sites—have changed over time.

Materials and Methods

Calculation of Normalized Substitution Rates

We calculated a list of ortholog triplets of mouse, rat, and human genes, using data downloaded from Ensembl (www.ensembl.org). The data were taken from human build NCBI35, mouse build NCBIM34, and rat build RGSC3.4, as of September 2005. We used the set of all known peptide sequences (pep.known-ccds for humans, and pep.known for mouse and rat) and determined reciprocal best BLAST (Altschul et al. 1990) matches (with a minimum significance of 1e-10) between pairs of protein sequences in each species. An ortholog triplet was accepted if the reciprocal best-hit relationship was satisfied for each of the three species pairings. CLUSTALW was used to calculate peptide alignments, which were then used as templates for alignments of cDNA, yielding 8487 aligned orthologous cDNA sequences.

The fourfold sites common to the three species were used to calculate substitution rates along each of the evolutionary branches: human to the rat-mouse ancestor and rat to mouse. Fourfold sites were defined to be third base positions for fourfold degenerate codons and in which mouse, rat, and human had identical bases in the two positions preceding and one position succeeding the fourfold site (to control for neighbor effects and tandem substitutions). To infer the substitution rate between human and rat-mouse ancestor, we used a parsimony assumption, restricting ourselves to the set of fourfold sites in which rat and mouse have the same base. For these sites, the common rat-mouse base was compared to the human base. For the mouse-rat lineage, we used all of the common fourfold sites. The overall percentage of nonmatching rat-mouse sites was $<p_{\text{rm}}> = 14.9\%$, and in the human-rodent common ancestor lineage, the percentage was $<p_{\text{rod-hu}}> = 27.5\%$.

For each gene the fraction of nonmatching fourfold sites was measured and normalized to correct for stochastic finite-size effects to yield rates $r_{\text{r-m}}$ and $r_{\text{rod-hu}}$ following a procedure described in Chuang and Li (2004). This involved defining an expected standard deviation σ based on the number of fourfold sites, N, in the gene, with $\sigma(N) \equiv \sqrt{<p>(1-<p>)/N}$. This standard deviation was then used to recalibrate the observed percentage of nonmatching sites, p, to a normalized rate $r \equiv (p - <p>)/\sigma(N)$, which yielded a rate distribution that was zero-centered and whose standard deviation was of order 1. On average, each gene had 166 fourfold sites and 141 sites identical between rat and mouse. For these values, rates within σ of the genome-wide frequency of nonmatching sites $<p>$ would be from 0.149 ± 0.028 for the rat-mouse lineage and 0.275 ± 0.038 for the rodent-human lineage. As a correction for base composition effects, the quantities $<p>$ and $\sigma(N)$, used to calculate $r_{\text{r-m}}$ and $r_{\text{rod-hu}}$, were recalculated based on the mouse or human base composition at the fourfold sites, respectively (Chuang and Li 2004). For example, for calculating the rat-mouse normalized rate for a given gene, we redefined

$$<p> = \sum_{\alpha=ACGT} f_\alpha <p>_\alpha,$$

where $<p>\alpha$ is the genome-wide average substitution rate of all mouse fourfold sites of type σ and f_α is the fraction of mouse fourfold sites of type σ in the gene of interest. A similar weighting was performed to redefine σ, but with the linear weighting on the variance, i.e.,

$$\sigma(N) = \sqrt{\frac{\sum_{\alpha=ACGT} f_\alpha <p>_\alpha (1-<p>_\alpha)}{N}}.$$

In practice this weighting had little effect on the rates. The weighted and unweighted rates had a correlation of 0.997 for the rat-mouse rate and a correlation of 0.989 for the rodent-human rate (Supplementary Data 1). It is possible that the restriction of certain analyses to bases in which rat and mouse coincide could bias those analyses toward bases that tend to be conserved across all species. However, the use of a z score mitigates this problem because it adjusts for any systematic bias.

In a preliminary analysis of the data, we found that correlations between the lineages could be significantly affected by genes with outlier rate values. To limit such effects from potentially spurious orthologs, we further restricted our rat-mouse-human orthologs to those in which the corresponding cDNA sequences also satisfied a three-way mutual best-hit relationship. This refined our dataset to 7462 high-confidence orthologs, which were used for all subsequent analyses.

For the dog analysis, 6250 1:1:1:1 orthologs were obtained using reciprocal best BLAST hits. The dog build was the BroadD1 build, which was downloaded from Ensembl. Normalized substitution rates were calculated using a procedure similar to that for the rat-mouse-human dataset. Rat-mouse rates and human-dog rates were each calculated using all aligned fourfold sites. Rates from the rat-mouse ancestor to the human-dog lineage were calculated by using only sites in which rat agreed with mouse and dog agreed with human. The fly data were downloaded from Flybase, using builds dmel-r4.0 and dpse-r1.03. Note that in each multispecies analysis, $<p>$ and z scores were recalculated based on the appropriate set of orthologs; this corrects for systematic biases in p that might be associated with that set of orthologs.

Other Substitution Rates

The parsimony criteria for the rodent-human branch were chosen to ensure that the rates $r_{\text{r-m}}$ and $r_{\text{rod-hu}}$ were calculated from independent data. However, we also considered a normalized rate based on the alternative rodent-human measure $(p_{\text{rat-human}} + p_{\text{mouse-human}} - p_{\text{rat-mouse}})/2$, which uses all the fourfold sites in each gene. We calculated the Pearson correlation of this measure with $r_{\text{rod-hu}}$. The correlation was 0.98, indicating that our results were not dependent on the restriction to only sites in which rat and mouse agree. For the Tamura-Nei, K80, REV, and CODEML maximum-likelihood mutation rate inferences, calculations were performed using PAML (Yang 1997).

Simulated Mouse-Rat-Human Sequences

PAML was also used to evolve simulated mouse-rat-human orthologous sequences. A "master set" of TN93 mutation

rates (without parsimony) was first inferred on each branch of the phylogeny, using the complete set of fourfold sites in all 7462 genes. The master rates correspond to the tree: ((Rat 1: 0.0857, Mouse 2: 0.082): 0.179, Human 3: 0.184). For each gene we then evolved a simulated set of mouse-rat-human sequences, with the same number of fourfold sites as the real gene. The simulated mutation rate along the (human- (rat-mouse ancestor)) branch was set equal to that of the real gene sequence, and the other rates were scaled proportionally using the rates in the "master set." The starting base composition of the ancestral sequence was chosen using the genome-wide average fourfold base composition of all three species: (T, C, A, G) = (0.222, 0.321, 0.202, 0.255). For example, for the ortholog triplet (ENSRNOP00000010032, ENSMUSP00000046233, ENSP00000330284), the inferred TN93 mutation rate was 0.963 along the branch connecting the human to the rodent ancestor. We therefore created a set of simulated sequences for this gene using the scaled tree ((Rat 1: 0.0857, Mouse 2: 0.082): 0.179, Human 3: 0.184) * 0.963/(0.179 + 0.184). After the simulated sequences were generated, rates were inferred as for the real dataset.

Genome-wide and Local Correlations

P values for the genome-wide and chromosomal Pearson correlations were calculated using a standard *t* test, as implemented in the Python STATS module.

For the local correlation analysis, on each human chromosome we calculated the Pearson correlation of $r_{\text{r-m}}$ and $r_{\text{rod-hu}}$ within a moving window of 15 genes (average width = 5.5 Mb). For alternatively spliced genes, we used only the longest version of the gene in any window. *P* values for the local correlation were assigned based on a bootstrapping procedure applied to each chromosome individually, which corrects for the fact that different chromosomes can have systematic biases in mutation rate (Castresana 2002). For a given chromosome, we randomly permuted the ordering of genes 1000 times, and for each permutation we measured the maximal value of the Pearson correlation, $c_{\max}$, over all windows. The observed distribution of $c_{\max}$ values was used to determine the significance of the correlation for any given window in the observed chromosome data, i.e., for correlation c, the p value of c was the fraction of the 1000 random permutations for which $c_{\max}$ was smaller than c. This method, which is based on $c_{\max}$ rather than on the complete distribution of all c values, was chosen to avoid biases caused by the fact that certain gene permutations yield many windows with strong correlation, whereas most permutations have no windows with strong correlation. This method has an expectation that if the observed chromosome data are random, only one window should have $p < 0.5$. To identify the genome-wide set of windows with significant correlation, we tabulated those for which the p value was less than 1/23, since there were 23 chromosomes to be considered. As described in the main text, we would expect only one window in all the chromosomes to meet our cutoff criterion by chance, but in actuality we observed a much larger number of windows.

Gene Ontology Analysis

For each GO category, we calculated a normalized substitution rate (z score) based on the substitution rates of all members of that category. The z score was defined to be

$$z \equiv \frac{\langle r \rangle_{GO} - \langle r \rangle_{all}}{\sigma_{all}/\sqrt{N_{GO}}}$$

where $\langle r \rangle_{\text{GO}}$ is the average substitution rate r for the genes in the GO category, $\langle r \rangle_{\text{all}}$ is the average r for all of the genes with Gene Ontology classifications, σ_{all} is the genewise standard deviation, and N_{GO} is the number of genes in the category. The p value for z was determined from the probability that a Gaussian-distributed variable takes on a value greater than or equal to z. We limited our analysis to the GO categories containing at least five genes, of which there were 875, and accordingly set a significance cutoff of $p < 1/875$ ($|z| \geq 3.23$), which corresponds to a conservative Bonferroni correction. The z scores z_{rm} and $z_{\text{rod-hu}}$ indicate the level of silent-site conservation in each branch of the phylogeny. A negative value indicates below-average substitution, while a positive value indicates high substitution.

We also calculated a z score for each category based on its contribution to the correlation of the two lineages. In this case, instead of using an r value for each gene, we used the quantity $x \equiv (r_{\text{r-m}} - \langle r_{\text{r-m}} \rangle) \times (r_{\text{rod-hu}} - \langle r_{\text{rod-hu}} \rangle)$. This quantity is proportional to each gene's contribution to the Pearson correlation, i.e.,

$$z_{corr} \equiv \frac{\langle x \rangle_{GO} - \langle x \rangle_{all}}{\sigma_{all}(x)/\sqrt{N_{GO}}},$$

where $\sigma_{\text{all}}(x)$ is the standard deviation of x over all genes.

To calculate whether any GO categories were overrepresented among the 670 genes in highly conserved local regions, we used hypergeometric statistics. For each GO category, we identified the number of genes in this set that were in the specified category. This number was compared to the number of genes in the category within the complete set of 7462 genes. The hypergeometric probability of having this many genes or more in the category was calculated exactly using Mathematica. This was supplemented

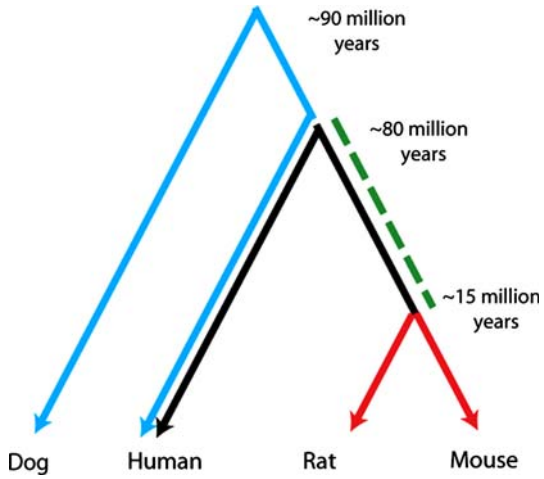

Fig. 1 Phylogeny of *Homo sapiens, Mus musculus, Rattus norvegicus*, and *Canis familiaris.* We calculated normalized synonymous substitution rates for reciprocal best-hit orthologs along several mammalian lineages: human-dog (blue), rat-mouse (red), human-rat/mouse ancestor (black), and the lineage connecting the human-dog branch to the rat-mouse ancestor (green dashes). The solid black line indicates the branch measured by the rate $r_{\text{rod-hu}}$ and the red line indicates the independent branch measured by the rate r_{rm}. Phylogeny and estimates of divergences times are from Springer et al. (2003)

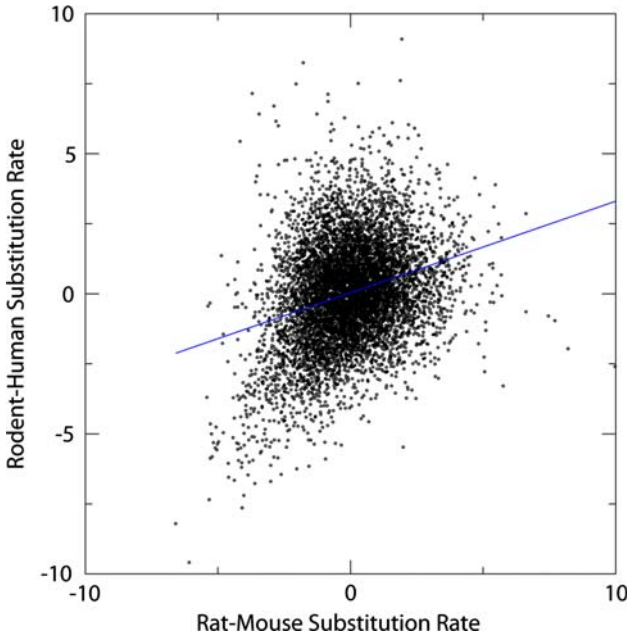

Fig. 2 Correlation of normalized synonymous substitution rates between the mouse-rat and human-rodent common ancestor lineages. The two rates $r_{\text{rod-hu}}$ and r_{rm} are significantly correlated (regression line is shown in blue), illustrating the repeatability of synonymous substitution rates. The effect is most prominent for several genes with low substitution rates in both branches of the phylogeny and is weaker at high substitution rates

by a parent-child union analysis performed with Ontologizer (Steffen et al. 2006) to correct for parent-child relationships in the Gene Ontology pseudohierarchy. Ontologizer was also used to assess the significance of finding multiple categories related to gene regulation among the 5% of all genes with the largest values of x, using a parent-child union analysis and a Bonferroni cutoff.

Results

Synonymous Substitution Rates Are Similar Across the Human-Rat-Mouse Phylogeny

We first measured synonymous site substitution rates within the rat-mouse-human phylogeny and assessed the repeatability of these rates across species. As shown in Fig. 1, humans diverged from rodents approximately 80 Myr ago, and rats and mice diverged approximately 15 Myr ago (Springer et al. 2003). Using the fourfold degenerate synonymous sites alignable in all three species, we first determined substitution rates in two evolutionary branches: human to the mouse-rat ancestor (black) and rat to mouse (red).

To determine repeatability, we measured the correlation of synonymous substitution rates in the two branches (Fig. 2). To guarantee independence of the data, the rates were inferred from distinct datasets according to a parsimony criterion. While the rat-mouse rates were calculated using all of the fourfold sites, the human- (mouse-rat ancestor) rates were inferred using only those sites for which rat and mouse shared the same base, and it was assumed that this was the ancestral base of the two species. The observed substitution rates were then mapped to z scores based on the observed number of substituted fourfold sites, the expected number of substituted fourfold sites, and expected variance, yielding rates r_{rm} and $r_{\text{rod-hu}}$ (see Methods). Because our z score is zero-centered and corrects for length effects (i.e., the greater expected variance in rates for genes with small numbers of fourfold sites), this method allows one to compare substitution rates from different lineages and different genes on a single scale. This z-score approach also corrects for systematic biases that may arise from choice of dataset. This method is also less prone to distortions from genes with substitution levels above saturation, which may have spuriously high inferred rates in maximum-likelihood approaches. To correct for base composition effects, the expected number of substituted sites and expected variance were linearly weighted according to the composition of the gene's fourfold sites and the values expected for each type of base (see Methods).

For a dataset of 7462 genes, the Pearson correlation was found to be 0.2899, which has a statistical significance of 10^{-144} according to a standard t test. We further analyzed

Table 1 Repeatability of silent substitution rates on each human chromosome

Dataset	Genes	Pearson correlation	p value
Chromosome 1	820	0.331	1.83E-22
Chromosome 2	471	0.358	1.13E-15
Chromosome 3	420	0.334	2.15E-12
Chromosome 4	268	0.285	2.19E-06
Chromosome 5	354	0.314	1.48E-09
Chromosome 6	379	0.342	7.67E-12
Chromosome 7	328	0.224	4.41E-05
Chromosome 8	263	0.316	1.65E-07
Chromosome 9	324	0.258	2.60E-06
Chromosome 10	296	0.238	3.41E-05
Chromosome 11	482	0.286	1.61E-10
Chromosome 12	405	0.368	1.86E-14
Chromosome 13	126	0.401	3.33E-06
Chromosome 14	252	0.325	1.30E-07
Chromosome 15	206	0.273	6.99E-05
Chromosome 16	328	0.133	1.58E-02
Chromosome 17	442	0.387	3.33E-17
Chromosome 18	91	0.277	7.79E-03
Chromosome 19	399	0.345	1.45E-12
Chromosome 20	267	0.220	2.86E-04
Chromosome 21	73	0.294	1.16E-02
Chromosome 22	169	0.102	1.88E-01
X Chromosome	235	0.234	2.98E-04
All chromosomes	7462	0.290	2.13E-144
All chromosomes, CpG Sites excluded	7462	0.302	6.78E-157

Pearson correlations between the substitution rates $r_{\text{rod-hu}}$ and r_{rm} are given for each human chromosome. 21/22 autosomes had significant correlation, and so did the X chromosome. Rates were strongly correlated on a genomic scale, with or without CpG sites. The total number of genes differs slightly from the sum from all chromosomes because a few genes did not have chromosome annotation. There were no eligible ortholog triplets on the human Y

the correlations of rates along each human chromosome, finding that the X chromosome and all but one of the autosomes had significantly correlated mutation rates (Table 1). The only chromosome with correlation p value larger than 1/23 was chromosome 22 ($p = 0.188$). The rest of the chromosomes had p values ranging from 0.012 (chromosome 21) to as low as 1.8×10^{-22} (chromosome 1), with a median of 2.2×10^{-6}. The correlation did not appear to be due to CpG effects. A strong correlation of the normalized rates remained even when fourfold sites overlapping a CpG in any of the species were ignored ($r = 0.3018$, $p = 6.8 \times 10^{-157}$), or under the more extreme restriction of ignoring any fourfold sites with a preceding C or a following G ($r = 0.1746$, $p = 3.7 \times 10^{-52}$). Rates were also strongly correlated when maximum-likelihood rates, rather than z scores, were used. PAML inferences of the K80 ($r = 0.1901$, $p = 1.1 \times 10^{-61}$) and TN93 ($r = 0.1741$, $p = 6.8 \times 10^{-52}$) substitution rates were both highly significant.

The correlations we observed were not dependent on the parsimony assumptions of the model. We verified this by first testing the relevance of restricting our analysis to only those sites in which rat and mouse agree. To do this, we calculated another z-score rate for the rodent-human branch using all fourfold sites; this alternative z score was based on the quantity $(p_{\text{rat-human}} + p_{\text{mouse-human}} - p_{\text{rat-mouse}})/2$, where the $p_{\alpha\beta}$ indicates the fraction of substituted sites between species α and β in the genes of interest. We found that the scores calculated from this measure and the original method were nearly indistinguishable (correlation of 0.98). The rat-mouse normalized rate was still strongly correlated with this rodent-human rate ($r = 0.1920$, $p = 7.3 \times 10^{-63}$), showing the irrelevance of the site restriction assumption.

As a second test of the parsimony assumptions, we performed TN93 and K80 maximum-likelihood rate inferences with the dataset expanded to the full set of fourfold sites and removing the assumption that sites in which rat and mouse agree indicate their ancestral base. We then inferred the substitution rates in each branch with PAML. Following this procedure, we found weaker correlations (TN93: $r = 0.0454$, $p = 2.9 \times 10^{-5}$; K80: $r = -0.0028$, $p = 0.8062$) than were observed for the original method. However, on manual inspection of the data, we found that this lack of correlation was strongly biased by a few outliers with extremely high substitution rates. These high rates are probably artifacts of the maximum-likelihood inference method, which produces large and unreliable rates when substitution is above saturation levels (note that a parsimony assumption tends to skew rates downward, mitigating these issues). Therefore, we removed all genes with an inferred substitution rate greater than unity in either branch (<4% of the genes) and recalculated the rate correlations. This procedure resulted in recapitulation of the strong correlation in the rates for both the TN93 measurements (7331 genes, $r = 0.2076$, $p = 3.3 \times 10^{-72}$) and the K80 measurements (7440 genes, $r = 0.2336$, $p = 9.6 \times 10^{-93}$), as well as in a CODEML codon-based rate analysis (7152 genes, $r = 0.3216$, $p = 1.0 \times 10^{-171}$). Thus, except for these outliers, the rate correlation is strong for these models as well. Because of this general robustness to different models, results described below are based on rates using the parsimony assumption, unless otherwise noted.

Similarity of Synonymous Substitution Is Consistent with Selection

Having established that synonymous substitution rates are correlated in these lineages, we then asked: Why does such

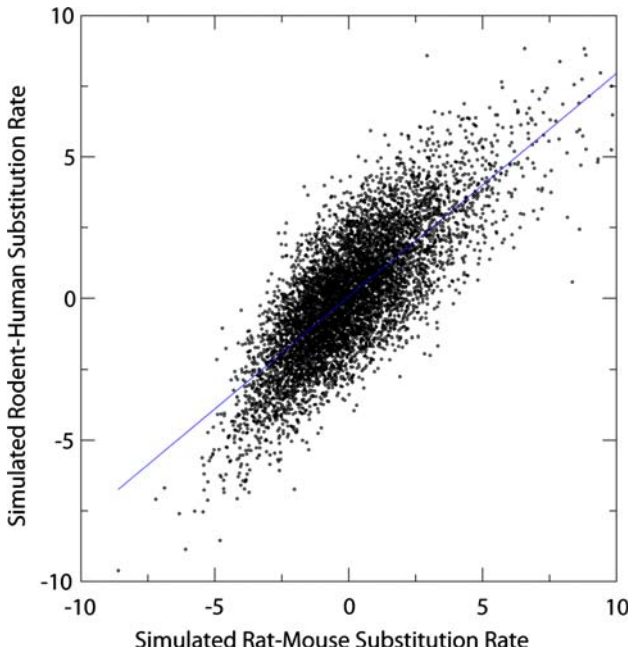

Fig. 3 Simulated sequences show no asymmetry in the repeatability at high and low substitution rates. We evolved a simulated set of mouse-rat-human sequences with the same distribution of fourfold sites and rodent-human mutation rates as the set of real genes. For each gene, the rates used to evolve the rat, mouse, and human sequences were given the same ratios, forcing the substitution rates to be repeatable across species. For this simulated dataset we then inferred the rates $r_{\text{rod-hu}}$ and r_{rm}, and these are plotted in the figure (regression line is shown in blue). There was no noticeable difference in the correlations at $r_{\text{rm}} < 0$ (Pearson correlation $r = 0.63$, 3960 genes, $p < 10^{-250}$) and $r_{\text{rm}} \geq 0$ ($r = 0.61$, 3502 genes, $p < 10^{-250}$), indicating that the asymmetry in Fig. 2 is not an artifact of the rate inference method

similarity exist? One possibility is that selection has acted on some synonymous sites, causing them to be conserved across multiple species. This effect would cause certain genes to have low substitution rates in each lineage. Another possible contributing factor is the persistence of neutral mutation patterns. It is well known that mammals have heterogeneous neutral mutation rates along their genomes, and the repeatability of substitution rates may simply indicate that the pattern of such mutation rates has not yet diverged in the different species. Recently, several groups have provided evidence for selection on synonymous sites (Chamary and Hurst 2004; Chamary et al. 2006; Hurst and Pal 2001; Schattner and Diekhans 2006; Smith and Hurst 1998; Xing and Lee 2005). We therefore examined our dataset to see if it was consistent with such claims of selection.

Similarity is stronger for genes with lower substitution rates

As can be seen qualitatively in Fig. 2, we observed that the correlation of rates is stronger for genes with lower rate scores, i.e., those with stronger sequence conservation. Still, there is significant correlation for other genes (when the 5% of genes with lowest r_{rm} are excluded, the correlation of r_{rm} and $r_{\text{rod-hu}}$ is 0.20, with $p = 6.3 \times 10^{-64}$). Using the average rate as the dividing point, we observed that the genes with $r_{\text{rm}} < 0$ have correlation $r = 0.33$ (3844 genes, $p = 1.2 \times 10^{-99}$), and the genes with $r_{\text{rm}} \geq 0$ have a smaller but still significant correlation of $r = 0.09$ (3618 genes, $p = 3.1 \times 10^{-8}$). This type of asymmetry between high and low substitution rates is consistent with purifying selection acting on some genes. For such genes, purifying selection will induce sequence conservation in the two lineages, leading to the strong correlation at $r_{\text{rm}} < 0$. For the remainder of genes, persistent neutral mutation processes will be the only relevant influence. Thus, the weaker correlation at high substitution rates is due to these substitutions largely reflecting the neutral mutation rate, without being masked by the effects of selection.

This asymmetry is not due to differences in the numbers of fourfold sites, N, in high- and low-rate genes. Our normalized rate method corrects explicitly for N dependence in the variance of rates, and the average N values for the $r_{\text{rm}} \geq 0$ genes (~ 172) and $r_{\text{rm}} < 0$ genes (~ 167) are similar.

Might the asymmetry be an artifact of some other aspect of our rate inference method? To test this, we created simulated orthologous rat-mouse-human sequences using the *evolver* program in PAML (see Methods). We chose mutation rates for evolving the sequences by first inferring a "master set" of TN93 mutation rates (without parsimony) on each branch of the phylogeny, using the complete set of 1,407,616 orthologous fourfold sites in all 7462 genes. For each gene we then evolved a simulated set of mouse-rat-human sequences, with the same number of fourfold sites as the real gene. In the sequence evolution procedure, the input mutation rate along the (human- (rat-mouse ancestor)) branch was set equal to that of the real gene sequence. The input mutation rates along the other branches were then scaled to be proportional to this rate, with the proportionality determined by rates in the "master set." This procedure provided a set of sequences with lengths and absolute substitution rates similar to the real set, but with repeatability of the underlying mutational process forced explicitly.

Using this set of simulated sequences, we tested whether the asymmetry between $r_{\text{rm}} < 0$ and $r_{\text{rm}} \geq 0$ was inherent to the method of calculating r_{rm} and $r_{\text{rod-hu}}$. We found that in the simulated set, there was no discernible difference in correlation between the $r_{\text{rm}} < 0$ set ($r = 0.63$, 3960 genes, $p < 10^{-250}$) and the $r_{\text{rm}} \geq 0$ set ($r = 0.61$, 3502 genes, $p < 10^{-250}$). This can be seen qualitatively in Fig. 3. The simulated data showed a strong overall correlation between r_{rm} and $r_{\text{rod-hu}}$ ($r = 0.77$, $p < 10^{-250}$); however, unlike the real data, the simulated data showed a strong correlation at

Table 2 Ten most correlated Gene Ontology categories in the rat-mouse-human phylogeny

GO ID	Genes	$z_{r\text{-}m}$	$z_{rod\text{-}hu}$	z_{corr}	Description
GO:0030286	6	1.999	2.301	8.866	Dynein complex
GO:0030374	10	−3.306	−4.089	8.752	Ligand-dependent nuclear receptor transcription coactivator activity
GO:0045944	5	−2.969	−4.001	8.035	Positive regulation of transcription from RNA polymerase II promoter
GO:0003723	195	−4.158	−5.637	7.417	RNA binding
GO:0005667	10	−2.856	−2.542	7.036	Transcription factor complex
GO:0001701	5	−3.025	−1.753	6.475	Embryonic development (sensu Mammalia)
GO:0016563	38	−3.429	−3.958	6.332	Transcriptional activator activity
GO:0003676	199	−0.480	−1.933	5.801	Nucleic acid binding
GO:0005634	1279	−5.959	−4.940	5.729	Nucleus
GO:0006378	5	−2.158	−2.115	5.403	mRNA polyadenylylation

Shown for each category are the number of genes, a z score for the silent substitution rate of the genes in the rat-mouse lineage ($z_{r\text{-}m}$), a z score for the silent substitution rate in the rodent-human lineage ($z_{rod\text{-}hu}$), and a z score (z_{corr}) for the contribution to the rate correlation between the lineages. Negative values of $z_{r\text{-}m}$ and $z_{rod\text{-}hu}$ indicate low substitution rates. Positive values of z_{corr} indicate similar substitution rates in the two lineages. The categories that contribute most to the correlation of rates tend to be related to gene regulation and have strong conservation of silent sites across the species

both high and low substitution rates. This indicates that the high/low asymmetry is not inherent to the method, and suggests that the stronger repeatability at low rates in the real set is due to selection.

The genes with ultraconserved coding sequences, defined by Schattner and Diekhans (2006) to be those containing a stretch of 60 codons with no more than five synonymous substitutions between mouse and human, or a stretch of 60 codons with at least three times as many nonsynonymous substitutions as synonymous ones, are illustrative of the connection between strong conservation and repeatability of rates. A number of such genes overlap with our dataset, and many have stretches that are almost indisputably due to selection, e.g., our ortholog triplet (ENSRNOP00000015831, ENSMUSP00000030315, ENSP00000257075) contains a run of 147 consecutive base pairs perfectly conserved across mouse, rat, and human. Overall, for the 171 ultraconserved genes in our dataset, we observed a much stronger Pearson correlation between r_{rm} and $r_{rod\text{-}hu}$ ($r = 0.51$, $p = 1.1 \times 10^{-12}$) than for even the $r_{rm} < 0$ set.

We also analyzed the GO categories with unusually similar substitution rates across species. We found that most had low substitution rates in each lineage, and many were related to gene regulation. For example, the ten gene families with the most similar rates across the phylogeny included the categories nucleic acid binding, transcription factor complex, transcriptional activator activity, and RNA binding, and all of these families had below-average substitution rates at their silent sites (Table 2). Similarity of rates was calculated via a z score z_{corr} from the values of $(r_{rm} - \langle r_{rm} \rangle) \times (r_{rod\text{-}hu} - \langle r_{rod\text{-}hu} \rangle)$ for the genes in each GO category (see Methods). Overall, we found 33 categories with significantly repeatable rates (minimum 5 genes, $|z| > 3.23$), and 25 of these have rodent-human substitution rates below the genome-wide average. The full set of categories is available in Supplementary Data 2. This bias toward regulatory genes is consistent with the results of Schattner and Diekhans (2006), who observed that genes with ultraconserved stretches of codons often had regulatory functions. This bias toward regulatory functions is not simply a consequence of the nonindependence of GO categories. An Ontologizer (Steffen et al. 2006) analysis, which accounts for parent-child inheritance relationships, of the 5% of genes with the highest values of $(r_{rm} - \langle r_{rm} \rangle) \times (r_{rod\text{-}hu} - \langle r_{rod\text{-}hu} \rangle)$ yielded 19 overrepresented categories, with four containing the annotation "regulation" (including the top two categories) and two containing the annotation "transcription."

Correlated substitution rates are often finely localized

Mutational environments have been measured to extend at least 10−15 Mb in mouse, rat, and human, a scale encompassing several tens of genes (Chuang and Li 2004; Gaffney and Keightley 2005). However, purifying selection should act on single genes. We therefore checked if the repeatability occurred at the scale of large mutational blocks or whether repeatability occurred on a finer scale. We observed a number of instances of strong local repeatability. An example is shown in Fig. 4 (top), which displays the region between 56 and 73 Mb on human chromosome 2, which contains 41 orthologs. Rates in the two lineages follow each other closely in trend and magnitude, even though substitution rates differ significantly from one gene to the next. We also found strong local correlations when genes were ordered according to their

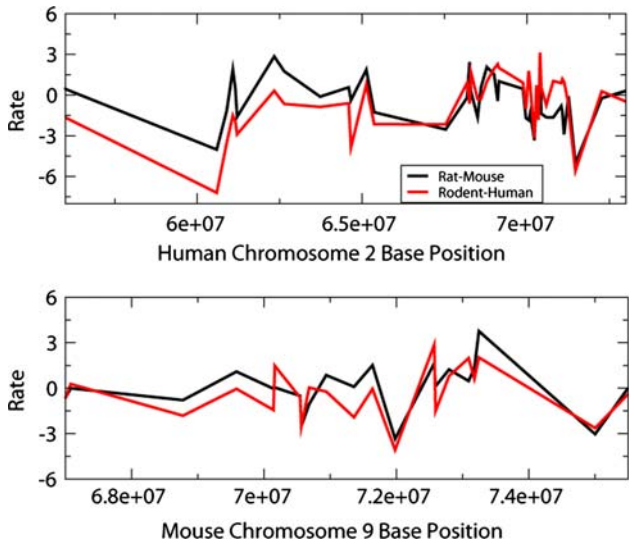

Fig. 4 A highly correlated region on human chromosome 2 (top) and another highly correlated region on mouse chromosome 9 (bottom). In these regions, rates in the two lineages follow each other closely in trend and magnitude, even though substitution rates differ significantly from one gene to the next. This suggests that these genes' synonymous substitution rates are a property of, at most, the few hundred kilobases around and including each gene

locations in the rat and mouse genomes. Figure 4 (bottom) shows an example from mouse chromosome 9. These observations are consistent with selection acting on the silent sites of the genes. In principle, this effect also could be caused by persistent neutral environments localized to a few hundred kilobases, which is not implausible given that some local neutral environments as fine as 100 kb have been reported in comparisons of human to the great apes (Reich et al. 2002). However, this second explanation would also require that neutral microenvironments be able to persist across divergences as far as human to the rodents, a question that has not yet been resolved.

To assess whether the highly correlated regions such as those shown in Fig. 4 could have arisen by chance, we performed a bootstrap analysis. We measured the correlations of rates in 15-gene windows along each human chromosome and compared these to values measured for 1000 random permutations of the gene ordering of the chromosome (see Methods). Eighteen of the 22 autosomes, as well as the X chromosome, had at least one 15-gene window with significant rate correlation. In total, we found 146 windows (each with $r > 0.71$) with significant correlations, containing a total of 670 distinct genes. Thus, it is not uncommon for substitution rates to be repeatable at a scale shorter than the large mutational blocks. It is also worth noting that the Gene Ontology category for which this set of genes is most biased is *GO:0006355 regulation of transcription, DNA-dependent* ($p = 0.002$, hypergeometric statistics), similar to the GO categories of genes with strong conservation and also suggested by Schattner and Diekhans to contain genes under selection. An Ontologizer analysis accounting for parent-child inheritance similarly gave *GO:0030528 transcription regulator activity* ($p = 0.015$) as one of the most overrepresented categories.

Changes in Synonymous Substitution Rates Across Species

By analyzing the repeatability of fourfold substitution rates, we have shown that some underlying synonymous selective pressures, and perhaps neutral processes, have persisted across the mouse-rat-human phylogeny. In general, such influences may change over time, which should cause more diverged branches to have less repeatable substitution rates. To explore how synonymous site influences have changed, we added dog to our set of species and calculated normalized synonymous substitution rates for 6250 reciprocal best-hit orthologs along three lineages in the phylogeny: human-dog (blue), rat-mouse (red), and the lineage connecting the human-dog lineage to the rat-mouse ancestor (green dashes). This phylogeny is shown in Fig. 1. Base composition corrections were applied based on the human sequence, the mouse sequence, and the human-dog ancestor, respectively. The human-dog ancestor to the rat-mouse ancestor branch was analyzed using the parsimony criterion, i.e., we used only sites in which human and dog agreed and in which rat and mouse agreed.

First, we found that fourfold substitution rates in the dog-human lineage were still correlated with those in the rat-mouse lineage ($r = 0.23$, $p = 10^{-73}$), as shown in Fig. 5. Thus, despite the approximately 65 million years since these two branches were last connected, some selective and/or neutral influences on the silent sites have persisted. It is interesting to note that the central (green dashes) branch is more similar to the other two (blue and red) branches than those two are to each other. This greater similarity of branches that are more closely related to each other suggests a gradual change in the synonymous site influences over time. In addition, we found that the central branch was more similar to the dog-human branch ($r = 0.41$, $p = 10^{-125}$) than to the rat-mouse branch ($r = 0.28$, $p = 10^{-117}$), suggesting that the rodent substitution patterns are nonancestral. These three rates are plotted versus one another in Fig. 6.

To verify the robustness of these findings, we also tested them via a maximum-likelihood rate inference without the parsimony condition. We inferred rates using the TN93 mutation model, excluding all rates greater than unity to remove the effect of outliers. We observed the same features as in the normalized rate analysis: all the correlations were significant; the more closely related branches had

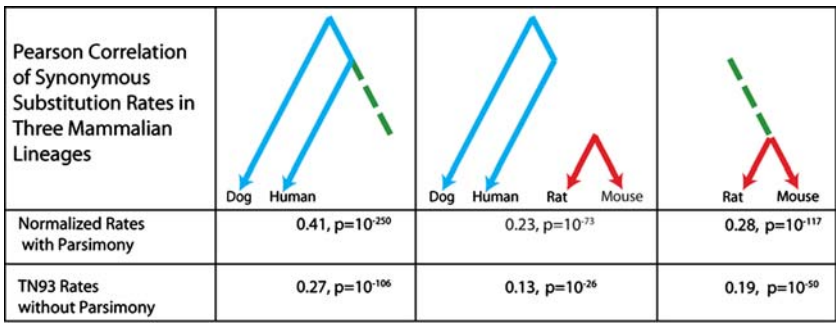

Fig. 5 Pearson correlations of synonymous substitution rates in three mammalian lineages. Fourfold substitution rates in all three lineages are significantly correlated with one another. The central (green dashes) branch is more similar to both of the other two (blue and red) branches than those two are to each other, suggesting a gradual change in the synonymous site pressures over time. The central branch is more similar to the human-dog branch than to the rat-mouse branch, suggesting that the rodent substitution patterns are nonancestral. These results are consistent for both the normalized method of inferring rates with parsimony and for the TN93 maximum-likelihood method without parsimony

more similar substitution rates than the distant branches; and the central branch was more similar to dog-human than to rat-mouse (Fig. 5).

We next tested if substitution rates were repeatable across even more distant clades, namely, mammals and flies. Given the apparent importance of selective effects to repeatability within the mammals, we reasoned that if repeatability were observed across mammals and flies, it would most likely be due to consistent silent site selective pressures. We identified a set of 1953 reciprocal best-blast-hit orthologs among rat, mouse, human, *Drosophila melanogaster*, and *Drosophila pseudobscura*. The fly protein and DNA sequences were downloaded from Flybase, and new normalized substitution rates were calculated based on sequence alignable among all five species. However, we found that the substitution rate along the lineage connecting the two flies was not repeatable with the mammalian lineages. No correlation was apparent in a visual inspection of the data, and the correlations of the fly rate with the rat-mouse rate ($r = 0.05$, 1953 genes, $p = 0.02$) and with the rodent-human rate ($r = 0.04$, $p = 0.07$) were both marginal. This was not a data size issue, as correlations between the rat-mouse and rodent-human lineages were still visible for these genes ($r = 0.16$, $p = 4.6 \times 10^{-13}$). Thus, the several hundred million years of divergence time from flies to mammals appears to have been sufficiently long that fly genes do not, in general, share the same synonymous substitution influences as their mammalian counterparts. This is despite the fact that flies also have regional mutation biases (J. Chuang, unpublished).

Persistence of Neutral Mutation Rates

As mentioned previously, we observed a statistically significant correlation ($r = 0.09$) between the rat-mouse and rodent-human rates for genes with $r_{rm} \geq 0$, which suggests that persistent regional neutral mutation processes also

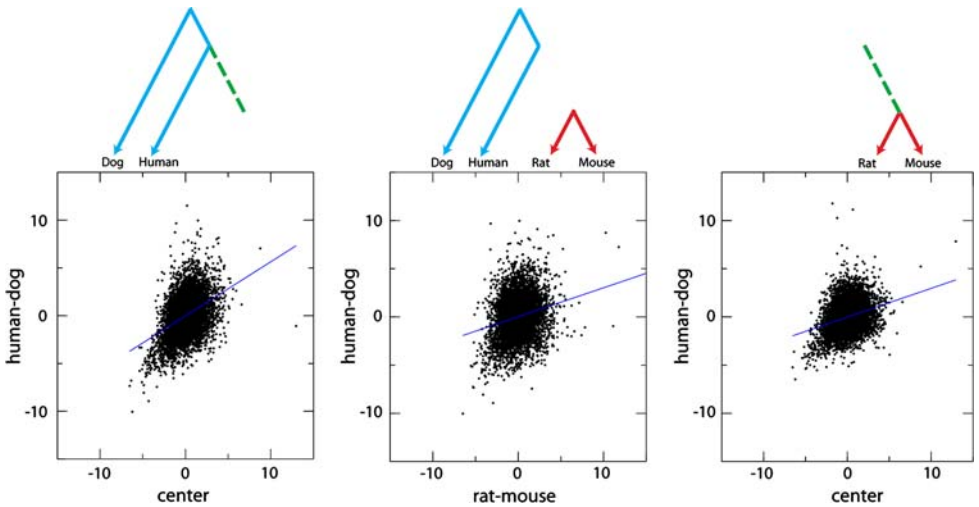

Fig. 6 Pairwise plots of synonymous substitution rates in three mammalian lineages. These graphs show the raw data used to calculate the correlations in Fig. 5. Regression lines are shown in blue

contribute to the repeatability of mammalian synonymous substitution rates. Given the small magnitude of the correlation, it seemed worthwhile to test the generality of this effect using the rat-mouse-human-dog phylogeny. In the rat-mouse-human-dog phylogeny, we again found significant substitution rate repeatability for genes with above-average substitution rate ($r_{\text{human-dog}} \geq 0$). The correlation of the human-dog lineage with the rat-mouse lineage was 0.14 (3168 genes, $p = 8.4 \times 10^{-15}$). This supports the idea that some regional neutral processes have persisted among the mammals.

Discussion

We have performed a large-scale analysis of synonymous sites in the mouse, rat, and human genomes, and our findings indicate that synonymous site substitution rates are repeatable in the lineages connecting the species. This is the main conclusion of our work. We found it to be robust to a number of different methods of inferring substitution rates, including: a *z*-score rate with corrections for base composition, a similar rate with CpG sites excluded, parsimonious and nonparsimonious approaches, and maximum-likelihood inferences via the K80 and TN93 models. We also observed repeatability of synonymous substitution rates in the rat-mouse-human-dog phylogeny, supporting the generality of our main conclusion. In the rat-mouse-human-dog phylogeny, the repeatability was stronger for branches in the phylogeny that are more closely related.

In an earlier study, Williams and Hurst (2002) reported a negligible correlation ($r^2 = 0.002$) in synonymous site substitution rates between the branches mouse-rat and human-cow (using 116 genes), arguing that previously observed correlations had been artifacts of particular mutation models. However, the Williams and Hurst conclusions are not generalizable beyond their small sample. In our 70× larger dataset, we found the observed correlations to be robust to multiple mutation models. Williams and Hurst observed the weakest correlations for maximum-likelihood rates inferred from fourfold sites using the Tamura-Nei model (TN93), but in our dataset the rates were very significantly correlated even for this model ($p = 6.8 \times 10^{-52}$). The correlations we observed were also robust to sampling of different subsets of the genes. Similar to the size of the Williams and Hurst study, we sampled 116 genes at a time and in each instance calculated the correlation of their TN93 rates. We repeated this procedure 10,000 times. In only 2.006% of these cases did we find a correlation less than or equal to what Williams and Hurst observed. Thus, their conclusions were likely the result of outlier effects from genes with high substitution rate or analysis of a nonrepresentative dataset.

Recently, a number of groups have analyzed mammalian transcripts for signs of purifying selection, with many positive indications (Chamary et al. 2006; Fairbrother et al. 2004; Kudla et al. 2006; Parmley et al. 2006; Xing and Lee 2005). Such selection appears to have been a palpable influence on the repeatability we observed. Our findings indicate that the genes that are likely to have been under purifying selection at their synonymous sites—most notably those with ultraconserved coding sequences, as well as genes with low substitution rates in general—tend to have more repeatable rates across the lineages. These genes with repeatable rates are often regulatory genes, consistent with the conclusions of Schattner and Diekhans (2006) that synonymous sites in regulatory genes are more likely to be functional. An interesting side note is that genes with low dN/dS also exhibit strong r_{rm} and $r_{\text{rod-hu}}$ correlation (5% of genes with lowest CODEML dN/dS values, $r = 0.44$, $p = 3.4 \times 10^{-19}$), suggesting that selection on amino acid sequence and selection on silent sites may occur simultaneously.

Aside from the repeatability caused by purifying selection, do other influences also play a role? Repeatability at above-average substitution rates is weaker than at low substitution rates, suggesting that neutral processes have not persisted as well as purifying selection pressures. Nevertheless, the genes with above-average substitution rates still have significantly correlated rates across species. This suggests that neutral processes have persisted to some extent across mouse, rat, human, and dog, which cover a span of several tens of millions of years. Of course, this is only an approximation to neutrality—a perfect dissection of the selective and neutral influences on these genes would require all selective pressures on synonymous sites to be known, and at the moment this goal is still distant.

Undoubtedly the neutral processes in some regions will have changed more than in others. For example, in an analysis of 1-Mb segments of DNA aligned across chimpanzee, human, mouse, and rat, the Chimpanzee Sequencing and Analysis Consortium reported a specific change in regional mutation rate from the hominids to the murids (the elevated mutation/recombination rates of the distal regions of hominid chromosomes were not observed in the rodents), providing an example of a change that could weaken the repeatability of neutral processes (Chimpanzee Sequencing Consortium 2005). Interestingly, the Chimpanzee Consortium reported an overall correlation of the substitution rates in these blocks; unfortunately, the bearing of this on neutral processes is not clear since these blocks contain genes and other functional elements.

Divergence time is an important consideration in whether neutral processes persist across species. Cooper et al. (2004) provided qualitative evidence that mouse and rat neutral mutation rates are similar to each other, and Smith et al. (2002) reported that chimpanzee and human rates are

correlated, with each group drawing their conclusions from analyses of blocks of genomic DNA purged of most functional sequence. Given the relatively short divergence times in each of these cases, it is reasonable that the regional mutation processes are similar; however, this begs the question of how quickly neutral mutation rates change. While our results suggest that regional neutral mutation processes can persist for approximately 65 Myr, we have also found that synonymous substitution rates are uncorrelated among fly and the mammals, which provides a loose upper bound of several hundred million years for rates to become uncorrelated. We expect that further repeatability analyses in other species will clarify how neutral mutation processes change, and will hopefully lead to a better understanding of what sequence features control regional mutation rates.

Acknowledgments This article was based upon work supported by the National Science Foundation under Postdoctoral Fellowship NSF-030698. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. JC thanks P. Arndt, C.S. Chin, S.L. Chen, J. Plotkin, and B. Tuch for comments on the manuscript. HL acknowledges support from NIH (grant GM 70808) and a David and Lucile Packard Fellowship.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391–1394

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc Natl Acad Sci U S A 88:5974–5978

Bustamante CD, Nielsen R, Hartl DL (2002) A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol Biol Evol 19:110–117

Castresana J (2002) Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. Nucleic Acids Res 30:1751–1756

Chamary JV, Hurst LD (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol Biol Evol 21:1014–1023

Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6:R75

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108

Chimpanzee Sequencing Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

Chuang JH, Li H (2004) Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS Biol 2:E29

Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, Sidow A (2004) Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res 14:539–548

Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2:E268

Gaffney DJ, Keightley PD (2005) The scale of mutational variation in the murid genome. Genome Res 15:1086–1094

Graur D, Li W-H (2000) Fundamentals of molecular evolution. Sinauer, Sunderland, MA

Hardison RC (2003) Comparative genomics. PLoS Biol 1:E58

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, Schwartz S, Furey TS, Whelan S, Goldman N, Smit A, Miller W, Chiaromonte F, Haussler D (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res 13:13–26

Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res 13:831–837

Hurst LD, Pal C (2001) Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet 17:62–65

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol 4:e180

Langley CH, Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. J Mol Evol 3:161–177

Lercher MJ, Chamary JV, Hurst LD (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res 14:1002–1013

Mouchiroud D, Gautier C, Bernardi G (1995) Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J Mol Evol 40:107–113

Parmley JL, Chamary JV, Hurst LD (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23:301–309

Rat Genome Sequencing Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32:135–142

Schattner P, Diekhans M (2006) Regions of extreme synonymous codon selection in mammalian genes. Nucleic Acids Res 34:1700–1710

Smith NG, Hurst LD (1998) Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. Genetics 150:823–833

Smith NG, Webster MT, Ellegren H (2002) Deterministic mutation rate variation in the human genome. Genome Res 12:1350–1356

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A 100:1056–1061

Steffen G, Sebastian B, Peter NR, Martin V (2006) An improved statistic for detecting over-represented gene ontology annotations in gene sets. Lecture Notes Comput Sci 3909:85–98

Williams EJ, Hurst LD (2002) Is the synonymous substitution rate in mammals gene-specific? Mol Biol Evol 19:1395–1398

Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory

motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434:338–345

Xing Y, Lee C (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc Natl Acad Sci U S A 102:13526–13531

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556